A Feature-Based Deisotoping Method For Tandem Mass Spectra

A Thesis Submitted to the

College of Graduate Studies and Research

in Partial Fulfillment of the Requirements

for the degree of Master of Science

in the Department of Biomedical Engineering

University of Saskatchewan

Saskatoon

By

Zheng Yuan

PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a Postgraduate degree from the University of Saskatchewan, I agree that the Libraries of this University may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by the professor or professors who supervised my thesis work or, in their absence, by the Head of the Department or the Dean of the College in which my thesis work was done. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to the University of Saskatchewan in any scholarly use which may be made of any material in my thesis.

Requests for permission to copy or to make other use of material in this thesis in whole or part should be addressed to:

Head of the Department of Biomedical Engineering

University of Saskatchewan

57 Campus Drive

Saskatoon, Saskatchewan

S7N 5A9

Canada

ABSTRACT

For high-resolution tandem mass spectra, the determination of monoisotopic masses of fragment ions plays a key role in the subsequent peptide and protein identification. It can directly influence the subsequent analysis of mass spectra including peptide determination and quantification. However, there are two difficulties during the process of detecting fragment ions: First, in some cases many real fragment ions have very low intensity and they can be removed as noise peaks by accident. Numerous noisy peaks in tandem mass spectra can cause either false negative or false positive fragment ions. Second, due to the existence of heavy isotopes in nature, more than one isotopic peak for each fragment ion is resolved in high-resolution tandem mass spectra. Though isotopic peaks can provide us with useful information, such as compound composition and charge states, they can increase the computational cost if peptide identification is done without removing them. In addition, isotopic peaks can overlap, which could result in wrong interpretation of masses of fragment ions.

In bottom-up proteomics, proteins are firstly cleaved into smaller peptides which are then used to be analyzed. Since tandem mass spectra of smaller peptides are easier than that of the intact proteins, bottom-up spectra are most often used in the identification of peptides and proteins. In this paper, to increase the accuracy of the peptide identification and reduce the complexity of tandem mass spectral analysis, we present a new algorithm for deisotoping the bottom-up spectra. Isotopic-cluster graphs are constructed to describe the relationship between all possible isotopic clusters. Based on the relationships in isotopic-cluster graphs each possible isotopic cluster is evaluated with a score function that is built by combining non-intensity and intensity features of fragment ions. The non-intensity features are used to prevent fragment ions with low intensity from being removed. Dynamic programming is adopted to find the paths with the highest score, which are presumably the most reliable isotopic clusters. Experimental results show that the average

Mascot scores and F-scores of identified peptides from spectra processed by our deisotoping method are greater than those by widely used YADA and MS-Deconv software.

Key words: tandem mass spectra, deisotoping, features, overlapping, isotopic-cluster graphs, dynamic programming.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisor, Professor Fang-Xiang Wu, for supporting my study and research during my staying in university. I really appreciate Dr. Wu that he introduced me to the field of deisotoping tandem mass spectra. Also I would like to thank Dr. Wu for his valuable comments and recommendations on my publications as well as this thesis, for his understanding and compassion regarding my personal situation. Without his supervision, it's impossible to complete this research.

I would like to thank the rest of my committee members: Dr. Tony Kusalik and Dr. Gopalan Selvaraj for their advice, suggestions, and comments on my research.

I thank my group members. They are Wenjun Lin, Jinhong Shi, Lizhi Liu, Xiaoyu Zhang, Yan Yan, Bolin Chen and Weiwei Fan. Thank you for providing me great help and a lot of advice.

Last but not the least, I would like to thank my family for giving me generous support and unbounded love.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CID: collision-induced dissociation

FDR: false discovery rate

TP: true positive

FP: false positive

FN: false negative

MS/MS: Tandem mass spectrometry

CHAPTER 1

INTRODUCTION AND PROBLEM DESCRIPTION

1.1 Introduction

1.1.1 Tandem mass spectrometry

Nowadays, tandem mass spectrometry (MS/MS) has obtained an important status in protein and peptide analysis. With the development of modern spectrometers, a large number of tandem mass spectra can be generated in a relative short time. It can be used in many fields, such as the acquisition of structure information and identification and qualitative analysis [1].

There are two mass spectrometers involved in MS/MS. The basic principle of MS/MS is illustrated in Figure 1.1: first, the ionization source brings the analyte into the gas phase and ionizes the analyte; second, the first mass analyzer (MS1) separates the ions according to their mass-to-charge ratio (m/z) and the ions are selected to be the precursor ions by MS1 based on the intensities of the ions; third, these precursor ions enter a collision chamber where they are fragmented into fragment ions; fourth, the m/z values and intensities of these fragment ions are measured by the second mass analyzer (MS2); finally, the detector generates signals. For the collision process, there are several fragmentation methods. However, the collision-induced dissociation (CID) is currently the most commonly used [2]. In the CID, the ions collide with a collision gas and some of the kinetic energy is converted into internal energy. Once the internal energy is enough to break chemical bonds, the decomposition of the precursor ions into smaller fragment ions occurs.

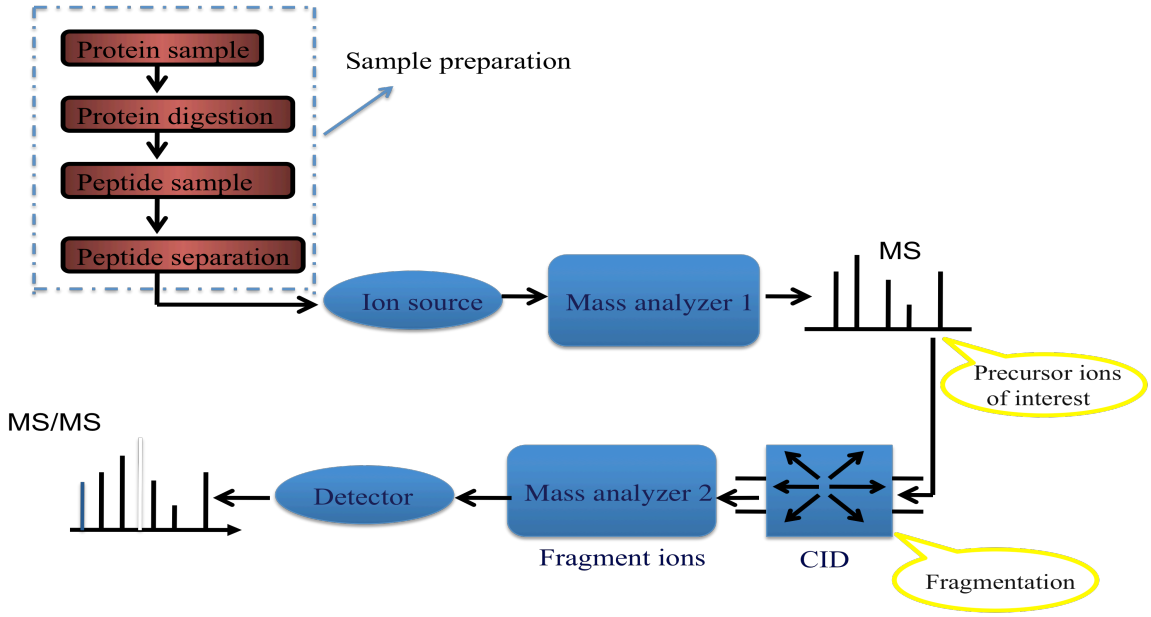


Figure 1.1 Basic principle of tandem mass spectrometry.

1.1.2 Proteins identification by tandem mass spectrometry

Proteins, also known as polypeptides, play an important role in all living organisms. The basic component of protein structure is that 20 different kinds of amino acids are linked by peptide bonds in a specific order determined by the sequence of a gene [3]. All amino acids are composed of one central carbon atom, one attached hydrogen atom, one amine group, one carboxylic acid group and one side chain. The differences between amino acids are determined by the differences of the side chains. The general structure of amino acid is showed in Figure 1.2:

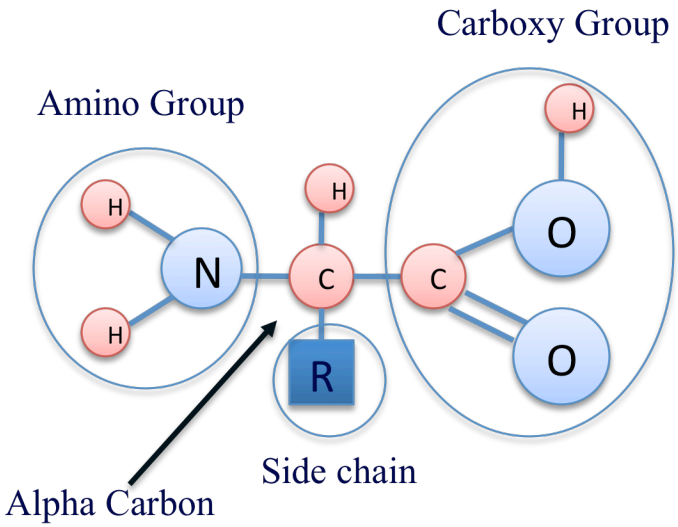


Figure 1.2 The basic structure of an amino acid.

Amino acids are combined together by peptide bonds to form a polypeptide chain. Figure 1.3 shows a short polypeptide chain. The end of the polypeptide terminated by an amino acid carboxyl group (-COOH) is called the C-terminal of the polypeptide while the start of the polypeptide terminated by an amino acid amine group (-NH2) is called the N-terminal.

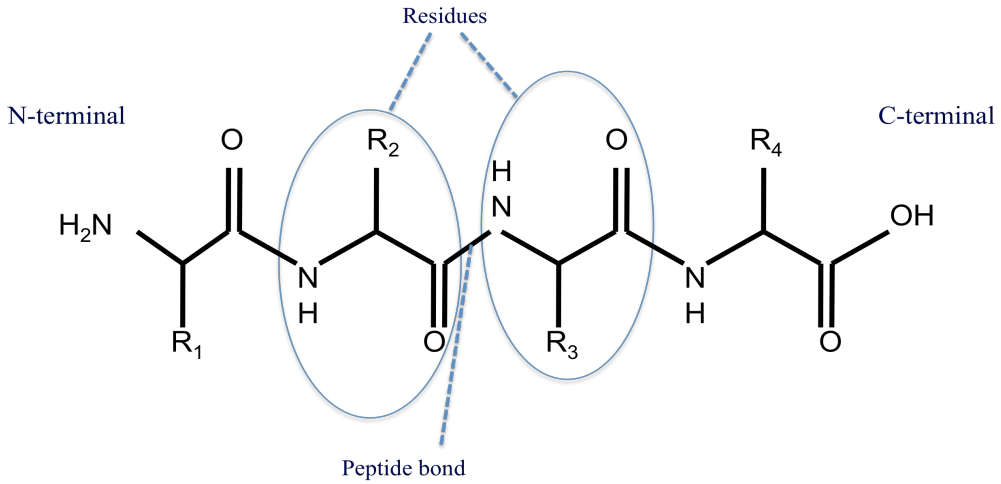


Figure 1.3 A structure of a short polypeptide chain. The left end of this peptide chain is N-terminal and the right end is C-terminal. In the circled part, there is an amino acid residue. Between these two, there is a peptide bond.

The determination of the sequence of peptides is critical in identifying proteins in biological samples. In the past, peptide analysis was performed by the Edman degradation procedure [1]. More recently, fragmentation by CID has been used for peptide identification. During sample preparation for MS/MS experiments, proteins are digested into peptides by enzymes. A precursor peptide ion is decomposed into fragments ions by CID. Figure 1.4 shows the fragmentation of a precursor peptide ion by CID. The breakages mainly occur in three different kinds of sites along the peptide backbone: CH-CO, CO-NH and NH-CH bonds. Thus, six likely types of fragment ions, also called backbone fragments, can be observed in a fragment ion spectrum: if the N-terminal of a fragment ion keeps a charge, this ion is classified as a a-, b-, c-ion; if the C-terminal of fragment ion keeps a charge, this ion is classified as a x-, y-, z-ion. In addition, when backbone fragments contain serine, lysine, threonine, aspartic acid or glutamic acid residues, water loss often occurs. The water molecule (-18.011Da) is usually lost from the side chains of these residues. When backbone fragments contain arginine, lysine, asparagine, or glutamine residues, ammonia loss often occurs. The ammonia molecule (NH3, -17.027Da) is usually lost from the side chains of these residues. Though other fragment types may be observed by further fragmentation, backbone fragments are predominantly observed in the MS/MS spectrum with low energy CID (<100eV). Thus, the most dominant real signal peaks in MS/MS spectrum represent b- and y- types of fragment ions and their derivatives [4].

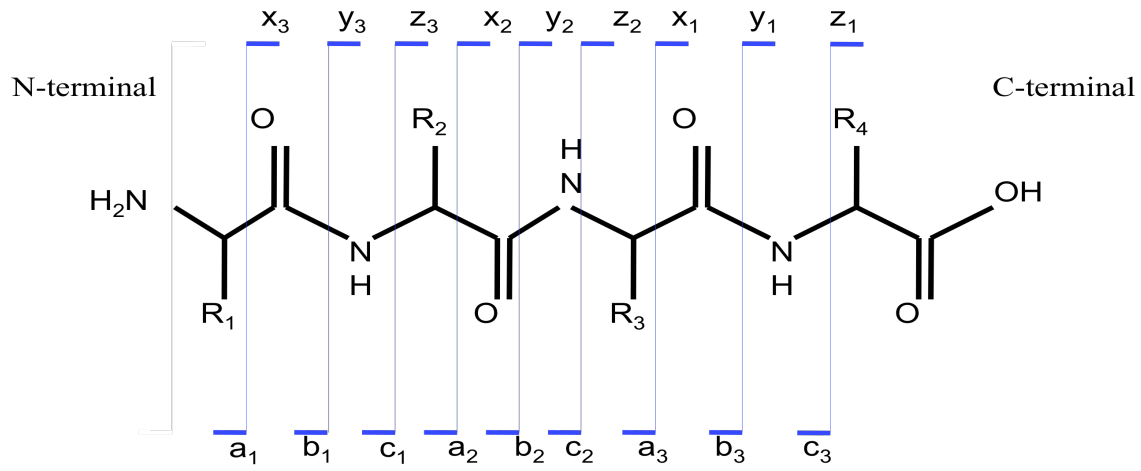


Figure 1.4 The fragmentation of a peptide by CID.

1.1.3 Isotopic pattern

As a matter of fact, there are many more peaks than those representing fragment ions in high resolution MS/MS. Besides noisy peaks from the electronic system and chemical contamination, more than one peak for each fragment ion exists in tandem mass spectra (Figure 1.5). Those peaks that represent the ions of the same element composition, but of different isotopic composition, are named as isotopic peaks. Such a group of isotopic peaks representing the same molecule is called an isotopic cluster. Generally, the first peak of an isotopic cluster is the monoisotopic peak. The monoisotopic peak represents the ion in which the composed elements are the most abundant naturally occurring stable isotopes. For peptide with small mass (<1800 Da), the most abundant peak is the monoisotopic peak [5]. The average mass difference is 1.003Da between the adjacent isotopic peaks in each isotopic cluster. Since x axis in the mass spectrum represents the m/z value, the space between the isotopic peaks is approximately 1.003/charge [4]. The presence of isotopic peaks is attributed to the existence of isotopes of the constituent elements ^{12}C, ^{13}C, ^{1}H, ^{2}H, ^{3}H, ^{14}N, ^{15}N, ^{16}O, ^{17}O, ^{18}O in peptides [6]. The natural abundance of these elements is already known. The most common element in peptides is carbon. In addition, the heavy isotope of carbon has much higher abundance than that of

heavy hydrogen, ^{2}H. Thus, the isotopic patterns of fragment ions are mainly affected by carbon [7]. The intensity ratio of adjacent peaks in each isotopic cluster can be predicted in terms of the natural abundance of the constituent elements and the element composition of this ion [8]. Since the fundamental difference between noise and ions is that ions have the isotopic pattern but not so with noises, the isotopic pattern can be used as a feature to distinguish real peaks from noisy peaks [9]. Moreover, isotopic patterns have been used in many aspects of MS/MS analysis, including establishing the composition of unknown molecules and determining fragmentation pathways. However, not like high-resolution mass spectra (the resolution ≥20000, commonly), the isotopic peaks cannot be resolved in mass spectra with low-resolution (the resolution ≤2000, commonly). For a peak with a given m/z value, the resolution R_m must be at least (m/z)*z, to resolve isotopic peaks [7].

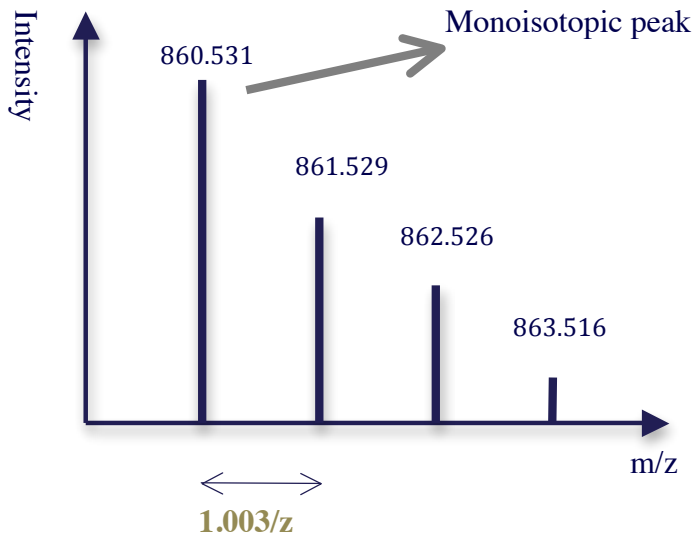


Figure 1.5 Isotopic peaks of an ion with m/z 860.531.

Due to the existence of isotopes and multiple charges, the isotopic cluster of one fragment ion may overlap with that of another fragment ion [10]. The isotopic intensity distribution of overlapping clusters will be different from that of a single one. Even for mass spectrometry with a high resolution, there are many overlapping isotopic cases in a tandem mass spectrum [11].

1.2 Problem statement

Since the fundamental data used for peptide identification in tandem mass spectra (MS/MS) is the mass value and charge states of fragment ions, their detection can directly affect the subsequent analysis of mass spectra including peptide identification and quantification [12]. However, there are two difficulties during the process of detecting fragment ions: First, in some cases real fragment ions have very low intensity and so are removed as noisy peaks by accident [13]. Numerous noisy peaks in tandem mass spectra can cause either false negative or false positive fragment ions. Second, due to the existence of heavy isotopes in nature, more than one isotopic peak for each fragment ion is resolved in high-resolution tandem mass spectra. Though isotopic peaks can provide us with useful information, such as compound composition and charge states, it will increase the computational cost if peptide identification is done without removing them. In addition, isotopic peaks can overlap, which could result in the wrong interpretation of masses of fragment ions. Thus, to increase the accuracy of the peptide identification and reduce the complexity of MS/MS analysis, an effective deisotoping method with the consideration of overlapping cases is necessary before peptide identification.

1.3 Objective and basic ideas

The goal of my research is to increase the accuracy of the peptide identification and reduce the complexity of MS/MS analysis by developing a better deisotoping method.

• Many existing deisotoping algorithms mainly take advantage of intensity information of the isotopic peaks to deisotope mass spectra. Actually, for MS/MS, there are other important non-intensity properties of fragment ions that also can assist in deisotoping. To avoid peaks of fragment ions with low intensities being removed as noisy peaks by accident, in my research we explored the properties of fragment ions in tandem mass spectra as follows: mass relationships, relative intensity ratios of adjacent isotopic peaks, the relationships between the isotopic distribution of fragment ions and so on.

• Moreover, most of deisotoping methods above take overlapping cases into account in order to detect the monoisotopic peaks in the overlapping cases. These methods identify the overlapping cases by subtracting the identified clusters, which could identify overlapping cases to some extent. If one isotopic cluster is incorrectly identified, the determination of the rest in a spectrum will be effected. In my study, we find a better way to describe and analyze isotopic clusters in the overlapping cases.

1.4 Thesis organization

In Chapter 1, we introduced the background of my research, the problem statement, the objective and basic ideas, and the outline of this thesis. In Chapter 2, we show a preliminary study for our deisotoping method. An experimental test is conducted and several problems of this method are shown. In Chapter 3, we present an improved algorithm for deisotoping the bottom-up spectra. In Chapter 4, to test our improved method: three experiments are conducted with the comparison with YADA and MS-Deconv software. The experimental results show that the average Mascot scores and F-scores of identified peptides from spectra processed by our deisotoping method are greater than those by YADA and MS-Deconv software. This indicates that our method performs better in deisotoping than those two pieces of software. In Chapter 5, the conclusion of our research is drawn. In addition, we give some suggestions for the improvement of our research in the future.

CHAPTER 2

A PRELIMINARY FEATURE-BASED DEISOTOPING METHOD FOR TANDEM MASS SPECTRA

2.1 Introduction

Many existing deisotoping algorithms [11-20, 23-29] have already been explored to detect the isotopic clusters of fragment ions. Some [11,14-19, 20] of these deisotoping methods are based on template matching, which means comparison of the theoretical isotopic distribution and the experimental isotopic distribution. The theoretical isotopic distribution can be estimated according to the monoisotopic mass of peptide ions [12, 14, 21-22]. If the observed signals match well with the theoretical isotopic distribution, these signals are considered as isotopic clusters and be subtracted from the spectrum. This procedure is repeated until no more possible isotopic clusters can be found. THRASH [14] is one of the most famous algorithms in the analysis of mass spectra. This algorithm has the following steps: determining noise intensity level; determining charge state by Fourier-Transform/Patterson techniques; estimating the composition of the peptide ions based on the average amino acid Averigine [23]; calculating theoretical isotopic distribution; and matching the theoretical isotopic distribution with the experimental one by the least-squares fitting to identify the monoisotopic peaks. Jaitly *et al.* [17] have developed Decon2LS that uses the similar components with THRASH but a different fitting scheme. The deisotoping speed of Decon2LS increases over THRASH. Unlike other algorithms [11,14-17,24], the OpenMS algorithm [18] is based on a combined two-dimensional model. One is the theoretical isotopic distribution that is calculated based on the approximate composition of peptides. The other is the elution curve estimated by a Gaussian distribution. However, the overlapping signal peaks in MS/MS can increase the computational cost. The major disadvantage of the template matching is that in case of

overlapping clusters, it is effective in identifying the isotopic clusters only based on the intensity information of the theoretical isotopic distribution and the experimental isotopic distribution. Once one isotopic envelop is incorrectly identified, the determination of the next isotopic envelop is easily impaired as in error propagation.

Sauelsson *et al.* [24] have proposed a quadratic programming deisotoping approach called Pepex in which observed spectra are modeled by a linear mixture model. Given the theoretical isotopic distribution and the observed isotopic distribution, the lowest number of peptides which can well explain the observed spectrum are determined by solving a quadratic programming problem. But in this method many parameters need to be optimized. In addition, only singly charged ions are considered. Renard *et al.* [25] have proposed a conceptually similar method NITICK which is also a statistical model based approach. The heart of this method is an iterative feature selection procedure which iterates over all relevant regions obtained from the raw spectra until a statistical termination criterion is met. An observed spectrum is modeled as a linear combination of expected isotopic profiles. The parameters of this model are estimated by minimizing the raw signal reconstruction error in a constrained regression problem (non-negative least squares). Moreover, to make this method more reliable, Renard *et al*. have improved the famous Senko's averagine model [23] to fractional averagine that is a mass error-free model. Compared with Pepex, there is no specific parameter optimization carried out in NITICK. However, the assumption of the error-free basis function in its non-negative least squares fit equation is violated even though they modified the averigine model.

Du *et al*. [26] have formulated the deisotoping issue as a statistical problem of the variable selection. This method selects the simplest model with the least number of isotopic clusters that can interpret MS/MS well. This method considers a spectrum as the intensity-weighted sum of isotopic clusters, where the intensity of each peak is expressed as a weighted sum of contributions from each cluster. Given an MS/MS, the intensities of all

potential isotopic clusters are considered as unknown variables in this method. The deisotoping problem becomes a variable selection problem that aims to find the most relevant and smallest subset of variables to interpret the spectrum well. The variables are selected by the LASSO method [27], which is an efficient procedure for automatically performing both the variable selection and the coefficient shrinkage for linear regression models. Further, errors in the theoretical isotopic distribution are also taken into account to avoid spurious isotopic clusters. This method avoids greedy feature selection as well. However, it's not reasonable to make the criterion "selecting the least number of isotopic clusters to explain the spectrum".

In contrast to the algorithms above, Zhang *et al*. [28] have developed a non-linear parametric model for one m/z interval. A Bayesian method is used to estimate the probabilities of the signal peak of an ion and the parameters of the model. For each signal peak, each charge state and isotopic position are considered. However, this method has not been implemented for peak detection at the peptide or fragment ion level. Sun *et al*. [29] have extended the method of Zhang *et al*. by developing a model for the whole spectrum considering the isotopic pattern and the charge state distributions. However, both methods only select signal peaks based on the intensity information of the observed spectra. Mcllwain *et al*. [30] have also used a Bayesian model to identify the isotopic distributions with a dynamic programming algorithm. This model is built to predict the probabilities of each potential isotopic distribution based on the number of isotopic peaks, the shapes of isotopic distributions, inter-distribution distance and intra-distribution distance. A dynamic programming algorithm is used to improve the sensitivity of the classifier and find an optimal sequence of isotopic distributions. However, overlapping cases are not taken into account in this method. The latter is a serious restriction limiting the method's applicability to complex mass spectra.

Carvalho *et al.* [20] have developed the freely available software YADA, which mainly deisotopes high-resolution middle-down spectra (the spectra for long polypeptides), but can process bottom-up mass spectra as well. This algorithm considers a peak as one possible isotopic peak when its intensity is larger than that of the previous peak and the space between peaks approximates 1/z. All possible isotopic peaks are stored in the same array. The normalized experimental isotopic distribution of each isotopic cluster is compared with the corresponding normalized averagine theoretical isotopic distribution. If they match, the isotopic cluster is determined. Liu *et al.* [31] have presented software MS-Deconv that can decharge and deisotope complex tandem mass spectra. This algorithm compares a set of theoretical isotopic distributions with the observed spectrum. And the set of peaks, which are matched with the theoretical isotopic distribution, is considered as the candidate isotopic cluster. The candidate isotopic cluster is scored based on the similarity between the experimental isotopic distribution and the theoretical isotopic distribution. Then, they construct a graph for the whole spectrum and select the isotopic clusters by searching for the heaviest path. However, both of these two pieces of software select the isotopic clusters based on the intensity information. Moreover, for MS-Deconv, since the graph is constructed after scoring the isotopic clusters, the relationships between isotopic clusters are not considered while scoring.

2.2 Preliminary Method

In this study, in order to solve the problems of the algorithms mentioned above, we present a new algorithm to detect the isotopic clusters of fragment ions and their monoisotopic masses in bottom-up spectra. Each peak in a spectrum is scored by a linear combination of four non-intensity features [32] that are based on the formation mechanism of peptide fragment ions from CID. The scores are used to remove noisy peaks and keep possible signal peaks in a spectrum. Then the experimental isotopic distributions of all possible isotopic clusters for each peak are compared with the corresponding theoretical

isotopic distributions. Moreover, several predominant overlapping cases are taken into account during the process of matching. The comparisons between the experimental and theoretical isotopic distributions are processed until there is no more isotopic cluster left in a spectrum. After the isotopic clusters determined, we output the monoisotopic peaks.

2.2.1 Select possible signal peaks in a spectrum

These following non-intensity features are used to distinguish signal peaks from noisy peaks in a spectrum. Some variables are defined before describing these features.

$$
\begin{aligned}
&diff1(x,y) = x - y \\
&diff2(x,y) = x - (y + M_H)/2 \\
&sum1(x,y) = x + y \\
&sum2(x,y) = x + (y + M_H)/2
\end{aligned}
$$

where x is the m/z value of one peak and y is the m/z value of another peak in the rest of the spectrum, respectively; M_H is the mass of a hydrogen atom. diff1 and sum1 considers that two fragment ions represented by x and y have the same charge state (z = 1, 2); diff2 and sum2 considers that the fragment ion represented by x is doubly charged and that represented by y is singly charged. Only singly charged fragment ions and doubly charged fragment ions are considered.

The first non-intensity feature (F_1) is the number of peaks y whose mass differences with another peak x approximate the residue mass of one of the twenty amino acids. Here, all methionine amino acids are considered to be sulfoxidized. Three pairs of amino acids are not distinguished because the mass of each pair is very close: isoleucine vs. leucine, glutamine vs. lysine, and sulfoxidized methionine vs. phenylalanine.

$$
\begin{aligned}
F_1 = |\{ y \mid & abs(diff1(x,y)) = M_{aa} + \theta \ \ or \\
& abs(diff1(x,y)) = M_{aa}/2 + \theta \ \ or \\
& abs(diff2(x,y)) = M_{aa}/2 + \theta \ \ or \\
& abs(diff2(y,x)) = M_{aa}/2 + \theta \} |
\end{aligned}
\qquad (2.1)
$$

where abs is the absolute value function; M_{aa} is the residue mass of one of twenty amino acids; |•| is the cardinality of a set; x and y can be any peaks in a spectrum. In this study, the fragment ion mass error tolerance θ is set as ±0.8 Da [32]. The first case of Equation (2.1) is for the case that peak x and peak y are both singly charged; the second case of Equation (2.1) describes peak x and peak y with double charges; the third case describes peak x with a single charge and peak y with double charges; the fourth case describes peak x with double charges and peak y with a single charge.

The second non-intensity feature (F_2) is the number of peaks y representing fragment ions that complement with fragment ion represented by a peak x. The sum of the mass of two complementary ions is equal to the mass of the precursor ion.

$$\begin{aligned} F_2 = |\{y \mid & sum1(x,y) = M + 2 \times M_H + \theta \ \ or \\ & sum1(x,y) = M/2 + 2 \times M_H + \theta \ \ or \\ & sum2(x,y) = M/2 + 2 \times M_H + \theta \ \ or \\ & sum2(y,x) = M/2 + 2 \times M_H + \theta\}| \end{aligned} \tag{2.2}$$

where M is the mass of the neutral precursor ion, and M_H is the mass of a hydrogen atom. x and y can be any peaks in a spectrum.

The third non-intensity feature (F_3) considers that the side chains of some amino acid residues of fragment ions can lose a water molecule (H_2O) or an ammonia molecule (NH_3). The number of peaks y whose mass differences with another peak x approximate the mass of a water molecule (H_2O) or an ammonia molecule (NH_3) is calculated.

$$\begin{aligned} F_3 = |\{y \mid & diff1(x,y) = M_{H_2O} \ \ or \ \ M_{NH_3} + \theta \ \ or \\ & diff1(x,y) = M_{H_2O}/2 \ \ or \ \ M_{NH_3}/2 + \theta \ \ or \\ & diff2(x,y) = M_{H_2O}/2 \ \ or \ \ M_{NH_3}/2 + \theta \ \ or \\ & -diff2(y,x) = M_{H_2O}/2 \ \ or \ \ M_{NH_3}/2 + \theta\}| \end{aligned} \tag{2.3}$$

where M_{H_2O} denotes the mass of a water molecule and M_{NH_3} gives the mass of an ammonia molecule; x and y can be any peaks in a spectrum.

The fourth non-intensity feature (F_4) considers two supportive ions, a ions and z ions which can be used to indicate the existence of the corresponding b ions and y ions. The number of peaks representing these kinds of supportive ions is determined. This feature measures the probability that that the differences between the masses of two peaks are the mass of a -CO group or an -NH group.

$$\begin{aligned} F_4 = |\{ y \mid & diff1(x,y) = M_{CO} \ \ or \ \ M_{NH} + \theta \ \ or \\ & diff1(x,y) = M_{CO}/2 \ \ or \ \ M_{NH}/2 + \theta \ \ or \\ & diff2(x,y) = M_{CO}/2 \ \ or \ \ M_{NH}/2 + \theta \ \ or \\ & -diff2(x,y) = M_{CO}/2 \ \ or \ \ M_{NH}/2 + \theta \}| \end{aligned} \tag{2.4}$$

where the mass of -CO is denoted by M_{CO} and the mass of -NH is denoted by M_{NH}; x and y can be any peaks in a spectrum.

The four features described above are linearly combined together in a score function to get the score of each peak [32].

$$Score = \omega_0 + \omega_1 \times F_1 + \omega_2 \times F_2 + \omega_3 \times F_3 + \omega_4 \times F_4 \tag{2.5}$$

where F_i (i=1,…,4) is the value of each feature, ω_i(i=0,...,4) are the coefficients. The bias ω_0 is set to 5 to ensure only a few peaks have negative score; $\omega_1$and $\omega_2$are set to 1.0; both ω_3 and $\omega_4$are set to 0.2. Ding *et al.* estimated these coefficients [32] based on the normalization method of the Sequest algorithm. This algorithm assigned a value of 50 to the b and y ions and 10 to the ions with the loss of a water molecule or an ammonia ion or a ions in a theoretical spectrum. The scaled values 1.0, 0.2 and 0.2 were assigned to the weights of F_1, F_3 and F_4. In this algorithm, the complementary ions were not considered. However, since both the first and second features were very important in the detection of fragment ions, the same value was assigned to the weights of these two features.

We assume a higher score for a peak means a higher probability that the peak is signal and not noisy peak. The intensities of peaks are adjusted by multiplying their original

intensities with the corresponding score. Thus after adjustment the intensities of signal peaks becomes higher and those of noise peaks become lower. Then a simple global threshold, which is 0.3 times the average of the adjusted peak intensities, is used to distinguish the real peaks from noisy peaks.

$$I_i' > threshold = 0.3 * \frac{\sum_{i=1}^{n} I_i'}{n} \tag{2.6}$$

where I_i' is the intensity of the peak after adjustment. Note that the intensities used in Section 2.2.2 are the original values. If the adjusted intensity of a peak is larger than the threshold, it is assumed to be a signal peak; otherwise, it is assumed to be a noisy peak.

2.2.2 Compare the experimental isotopic distribution with theoretical isotopic distribution

After roughly selecting the isotopic peaks, possible isotopic clusters for each peak that is retained are sought in a spectrum based on intensity feature of fragment ions. Searching starts from the peak P_0 with the lowest m/z value in a spectrum. Singly charged fragment ions and doubly charged fragment ions are considered in this method. The range of the number of isotopic peaks for one possible isotopic cluster is from 3 to 4; for one isotopic cluster, the spaces 1.003/z (z=1, 2) between each pair of adjacent isotopic peaks are approximately the same with the error tolerance 0.01 [9].

For each possible isotopic cluster, we compare its experimental isotopic distribution with the corresponding theoretical isotopic distribution.

$$\frac{|E_i - T_i|}{T_i} < threshold = 0.3 \tag{2.7}$$

where i is the order of a peak in one isotopic cluster; E_i is the experimental intensity of peak i; , T_i is the theoretical intensity of peak i.

The theoretical isotopic distribution of one fragment ion can be predicted, since the intensity ratio of the adjacent peaks in one isotopic cluster depends on the natural abundance of the composed elements and the elemental composition of this ion. Many methods of the calculation of theoretical isotopic distribution exist. One of the most well known methods is an average amino acid averigine developed by Michael W. Senko [23] from the PIR protein database. The molecular formula of averagine is $C_{4.9384}H_{7.7583}N_{1.3577}O_{1.4773}S_{0.0417}$ with molecular mass 111.1254 Da. Assuming that a particular molecular mass is known, the number of averigine units that a molecular formula has can be calculated. Then, the element composition of this molecule can be acquired. Also, the relative natural abundance of each element C, H, N and O is already known. Based on the information above, the theoretical isotopic distribution of an ion with a particular mass can be predicted. Actually, given the estimated elemental composition of a molecule, many existing software packages can compute its theoretical isotopic distribution. Here, in my method, it's done by the averigine method [23] with isotopicdist function in MATLAB.

2.3 Experimental test

To test this deisotoping method we employ MS/MS dataset A from Klammer *et al.* [33] that consists of 1208 high-confidence peptide-spectrum matches. This dataset was produced from a ThermoFinnigan LTQ-Orbitrap mass spectrometer. A sample from *Escherichia Coli* was used after being reduced, carbamidomethylated and digested with trypsin and without acid-labile detergent (RapiGest, Waters Corp., Milford, MA). The resulting peptides were analyzed by μLC-MS/MS, yielding a total of 112329 spectra. Of them, 1208 high-confidence peptide-spectrum matches generated by existing algorithms [34-36] were selected. The thresholds for getting those high-confidence peptide-spectrum

matches were set with a false discovery rate (FDR) of 1%. The charge range of spectra is from 1 to 2 while the mass range of spectra is less than 2000 Da.

2.4 Results

To evaluate the performance of this method, it was compared with YADA software [20] which can deisotope and decharge high-resolution middle-down spectra and bottom-up spectra. YADA is a freely available deisotoping and decharging tool that can process data in MS1 and MS2 format. The MS1 and MS2 file format is used to record MS/MS spectra. The results from an on-line Mascot [37] search were used to interpret the dataset processed by our deisotoping method and YADA. The parameters for the Mascot searching are listed in Table 2.1. The cysteine residues were set to be carboxamidomethyled as a fixed modification and methionine residues were set to be oxidized as a variable modification. All the searches used the SWISS-PROT database with one missed trypsin cleavage allowed. The tolerance for the peptide mass was ±1.2 Da and for the fragment mass was ±0.6 Da. In this study, the peptides are considered to be interpreted by Mascot searching engine within the false discovery rate (FDR) of 1%.

Table 2.1. The parameters for Mascot search

Enzyme	trypsin
Maximal missed cleavages	1
Fixed modifications	Carbamidomethyl
Variable modifications	Oxidation (M)
Peptide mass tolerance	±1.2 Da
MS/MS mass tolerance	±0.6 Da
Peptide charges	+1, +2, +3
Mass values	monoisotopic

It was assumed that the more peptides and proteins interpreted by Mascot, the better the effect of the deisotoping method was. Table 2.2 shows that the number of interpreted peptides and proteins in the raw data (dataset A), the processed data by YADA and that by our method. In total, 124, 181, 103 interpreted proteins are identified while 198, 273, 155 interpreted peptides are interpreted from the raw dataset, YADA method and our method, respectively. The results show that both the number of identified peptides and proteins from the dataset processed by our method is lower than that from the raw dataset and the data processed by YADA. They indicate that not only does YADA have a better effect on the Mascot search than our method, but also that the peptide and protein identification is not improved by our method. Since the results of our method are not good and we noticed some problems existed in this method, we need to improve it instead of doing more comparisons.

Table 2.2 Numbers of peptides and proteins identified by Mascot from dataset A (1208 spectra) processed by our method and YADA.

	Data processed by raw dataset	Data processed by YADA	Data processed by our method
proteins	124	181	103
peptides	198	273	155

2.5 Discussion

This method performs not as well as expected in deisotoping due to several problems as follows:

Firstly, the non-intensity features and intensity features of fragment ions work separately. The non-intensity features just work in roughly distinguishing the noisy peaks from the signal peaks. Only the intensity features work in finally determining the monoisotopic

peaks of fragment ions. The fragment ions with low abundance may be removed by mistakes as noisy peaks.

Secondly, each possible isotopic cluster is analyzed without considering the relationship with others. Once one isotopic cluster is incorrectly identified, the determination of the subsequent isotopic cluster will be easily impaired as in error propagation.

Thirdly, the number of isotopic peaks of each fragment ion in this preliminary method is set to 3 and 4. However, since the third and fourth isotopic peaks of each fragment ion with small masses (<600 Da) have very weak intensities that are easily removed or not easily observed, the limitation in the preliminary method may lead to the loss of a lot of fragment ions with small masses.

CHAPTER 3

AN IMPROVED FEATURE-BASED DEISOTOPING METHOD FOR TANDEM MASS SPECTRA

3.1 Improved Method

To solve the problems of our method previously presented in Chapter 2, an improved feature-based deisotoping method for tandem mass spectra is introduced.

Considering the complex overlapping cases, isotopic-cluster graphs are constructed to describe the relationship between possible isotopic clusters. Non-intensity properties [38] of fragment ions are explored in order to avoid removing those real fragment ions with very low intensities. They are combined with intensity properties of fragment ions in a score function. According to the relationship between isotopic clusters provided by isotopic cluster graphs, each candidate isotopic cluster is given a score based on the score function. Dynamic programming is adopted to find the path with the highest score as the optimal arrangement of isotopic clusters with the highest reliability. Moreover, the range of the number of isotopic peaks for each fragment ion is set from 2 to 4 instead of from 3 to 4. Triply charged fragment ions are considered besides singly and doubly charged fragment ions. To test our method, three experiments are conducted and compared with two pieces of software, YADA and MS-Deconv.

Our improved method for deisotoping is composed of four parts: searching for all possible isotopic clusters, constructing isotopic cluster graphs, scoring all possible isotopic clusters and searching for the path with the highest score. The first part aims to find all possible isotopic clusters. The second part describes the relationship between possible isotopic clusters. The third part assesses each possible isotopic cluster based on the assumed relationship. The goal of the fourth part is to determine the most likely arrangement of isotopic clusters.

3.1.1 Searching for possible isotopic clusters

Searching starts from the peak with the lowest m/z value in a spectrum. Firstly, all possible sets of isotopic peaks are determined based on three criteria as follows: each possible set (example shown in Figure 3.1) is composed of several peaks; the number of peaks in each set is not less than 2; the space between any pair of adjacent isotopic peaks in each set is 1.003/z (z=1, 2, 3) with an error tolerance 0.01. The starting peak P_s of each set is the first peak which is followed by one peak with the interval 1.003/z (z=1, 2, 3) between them; the ending peak P_e of each set is the last one which follows one peak with the interval 1.003/z (z=1, 2, 3) between them. For example, in Figure 3.1, set A consists of five peaks from peak P_s to peak P_e. The space between five adjacent peaks is 0.33 (≈1.003/z, z=3), 1 (≈1.003/z, z=1), 0.5 (≈1.003/z, z=2) and 0.5 (≈1.003/z, z=2).

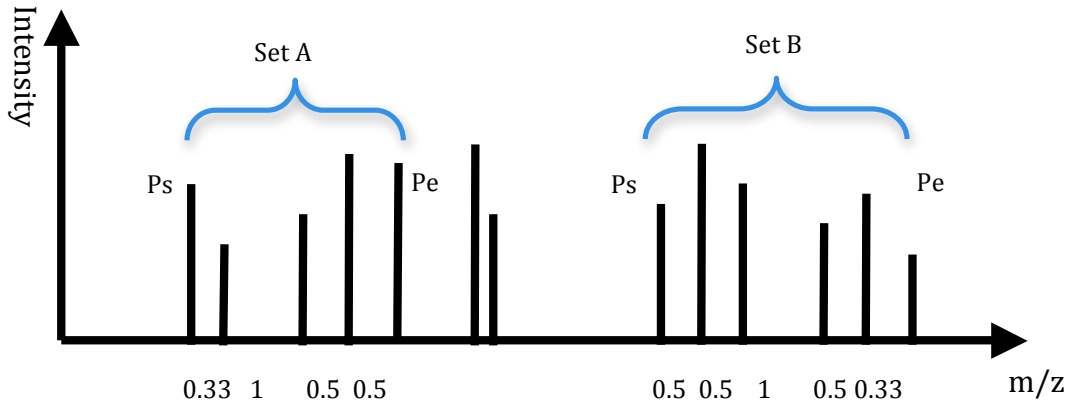


Figure 3.1. Sets of possible isotopic peaks. Note that those numbers on the x axis represent the distances between the adjacent peaks.

Secondly, in each possible set of isotopic peaks, we searched for all candidate isotopic clusters (shown in Figure 3.2). Each candidate isotopic cluster search follows two criteria: the range of the number of isotopic peaks for one possible isotopic cluster is from 2 to 4; for one isotopic cluster, the spaces 1.003/z (z=1, 2, 3) between each pair of adjacent isotopic peaks are the same within an error tolerance of 0.01. In Figure 3.2, the set includes six peaks. Isotopic cluster A and isotopic cluster B are two of possible isotopic

clusters in the same set. The space between any pair of adjacent peaks in isotopic cluster A is 0.5 (≈1.003/z, z=2). Isotopic cluster B is composed of three peaks of which any pair of adjacent peaks has the same interval 1 (≈ 1.003/z, z=1).

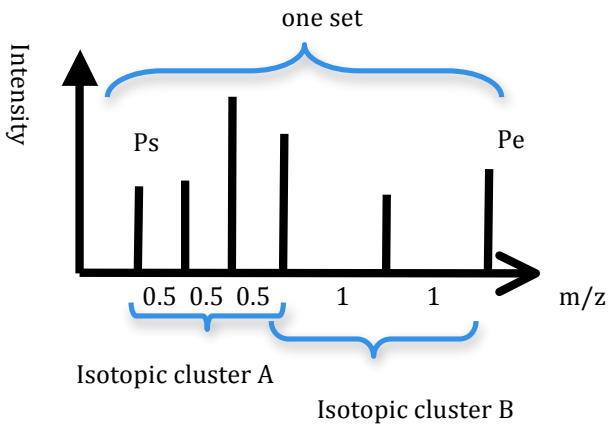


Figure 3.2 Possible isotopic clusters in one set

While searching for possible isotopic clusters, several predominant overlapping cases are taken into account. One situation is overlapping cases without sharing peaks (shown in Figure 3.3). Sets A and B, each of which includes five peaks P_0~P_4, are two examples. In Figure 3.3.A, one fragment ion is represented by an isotopic cluster composed of P_1 and P_3. The other isotopic cluster, composed of P_0, P_2 and P_4, represents the other fragment ion. There are no shared peaks in these two isotopic clusters. In Figure 3.3.B, both P_1 and P_3 are noisy peaks. An isotopic cluster composed of P_0, P_2 and P_4 represents one fragment ion.

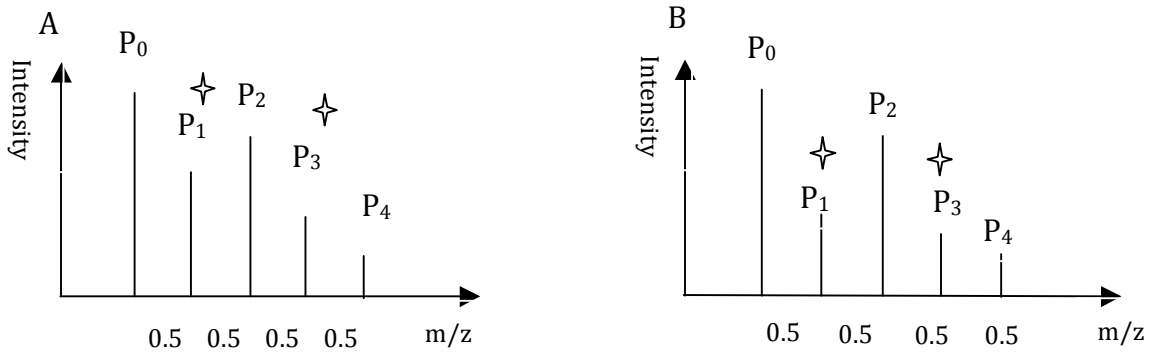


Figure 3.3. Cases without sharing peaks

The other situation is overlapping cases with sharing peaks (shown in Figure 3.4). In Figure 3.4.A, one fragment ion with a single charge is represented by an isotopic cluster composed of P_0, P_1 and P_2. The other fragment ion with a single charge is represented by a different isotopic cluster composed of P_1, P_2 and P_3. The overlap occurs at P_1 and P_2. In Figure 3.4.B, two isotopic clusters represent two singly charged fragment ions. One is composed of peaks P_0, P_1 and P_2 while the other is composed of peaks P_2, P_3. The overlap takes place in peak P_2. In Figure 3.4.C, one fragment ion, represented by the isotopic cluster composed of P_0, P_1 and P_2, is doubly charged. The other fragment ion, represented by the isotopic cluster composed of P_2 and P_3, is singly charged. P_2 is the overlapping peak. In Figure 3.4.D, one fragment ion, represented by the isotopic cluster composed of P_0, P_1 and P_2, is doubly charged. The other fragment ion, represented by the isotopic cluster composed of P_1, P_3 and P_4, is singly charged. P_1 is the overlapping peak.

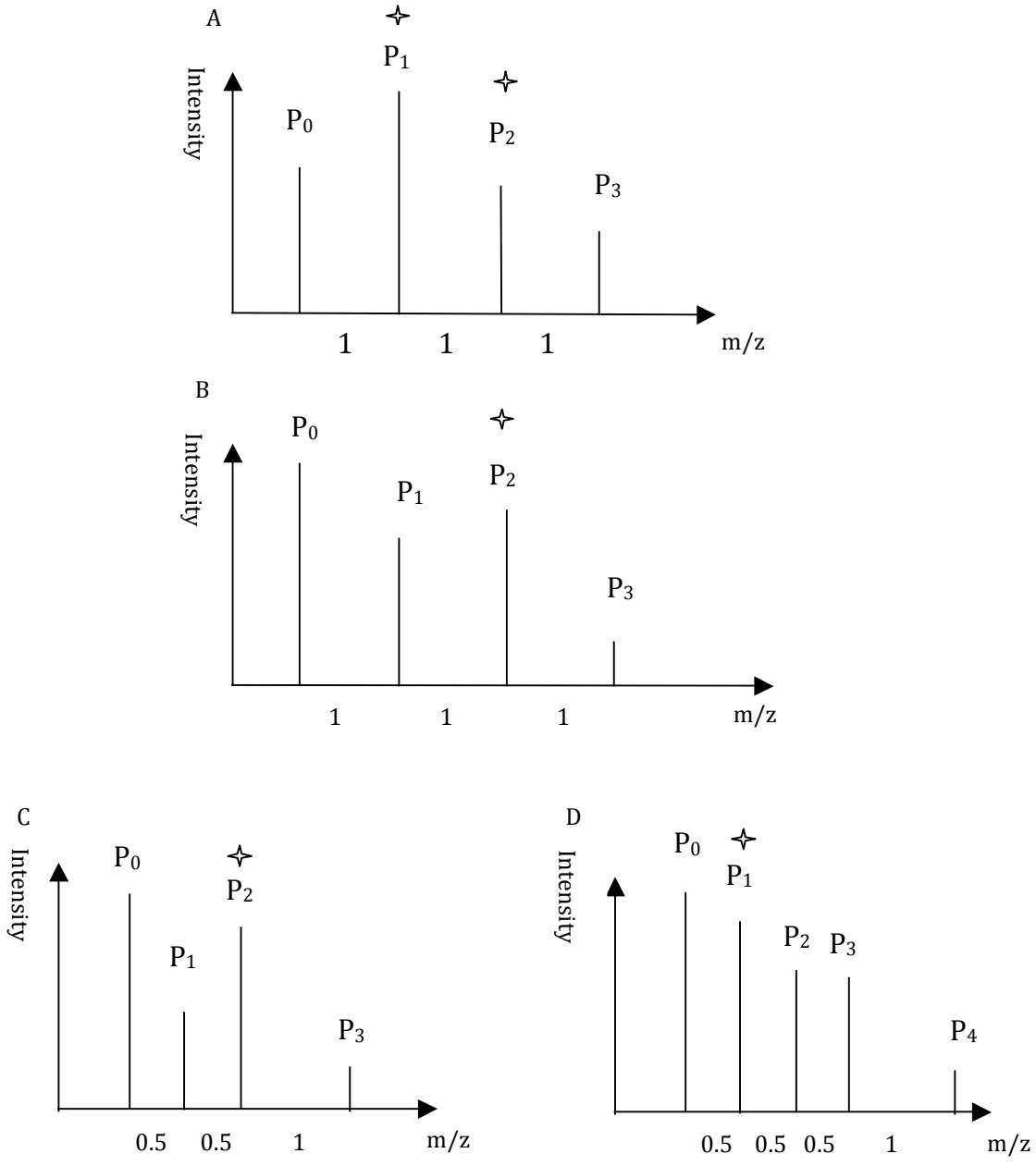


Figure 3.4 Overlapping cases with sharing peaks

3.1.2 Constructing Isotopic-cluster graphs

Graphs are often employed to model many bioinformatics problems [39]. New methods can sometimes be developed by modeling a problem as a graph. In addition, a solution of a problem sometimes can be directly obtained from graph theory. Problems involving relationships between objects, or concepts can be solved using graphs.

For each set in the spectrum, an isotopic-cluster graph is constructed to describe the predicted relationships among all possible isotopic clusters. Here, the relationship refers to whether or not two connected isotopic clusters overlap and how they overlap. Figure 3.5 illustrates how edges in an isotopic-cluster graph are expected to connect the possible isotopic clusters. Here, "connect" means that one cluster immediately succeeds the other in a graph. The source vertex in an isotopic-cluster graph is defined as the starting position of all paths while the sink vertex is defined as the ending position. A vertex in an isotopic-cluster graph is defined as one possible isotopic cluster generated by one possible fragment ion. Two types of edges are constructed in an isotopic-cluster graph: red edges represent that two adjacent isotopic clusters overlap; black edges represent two adjacent isotopic clusters connecting without overlapping. A black edge connects two isotopic clusters if the m/z value of the last peak in the source cluster (at the tail of the edge) is less than the m/z value of the first peak in the determination cluster (at the head of the edge). For example, in this figure, the m/z values of the peaks in each vertex are increasing. The m/z of the last peak P_3 in vertex A is less than that of the first peak P_4 in vertex E. Thus, the black edge is used to connect these two vertices. A red edge connects two isotopic clusters if they satisfy both the following rules: a) The m/z value of the first peak of the tail (source) of an edge is smaller than that of the head of this edge. Moreover, the m/z value of the last peak of the tail (source) of this edge is larger than that of the first peak of the head of this edge. b) If the number of isotopic peaks of the tail of an edge is 2, the second isotopic peak of this tail overlaps with the first isotopic peak of the cluster at the head of this edge. For example, the number of isotopic peaks in vertex G is two. The connected vertices G and F overlap at peak P_6; If the number of isotopic peaks of the tail of an edge is 3 and has one shared peak with the cluster at the head of this edge, the second or third isotopic peak of the tail overlaps with the first isotopic peak of the head. For example, the connected vertices E and F have one shared peak. The overlapping occurs at peak P_6; If the number of isotopic peaks of the tail of an edge is 3 and there are

two shared peaks with the cluster at the head of this edge, the second and third isotopic peaks of the tail, respectively, overlap with the first and second isotopic peaks of the head. For example, the connected vertex E and G have two shared peaks. The second and third isotopic peaks of E, respectively, overlap with the first and second isotopic peaks of G.

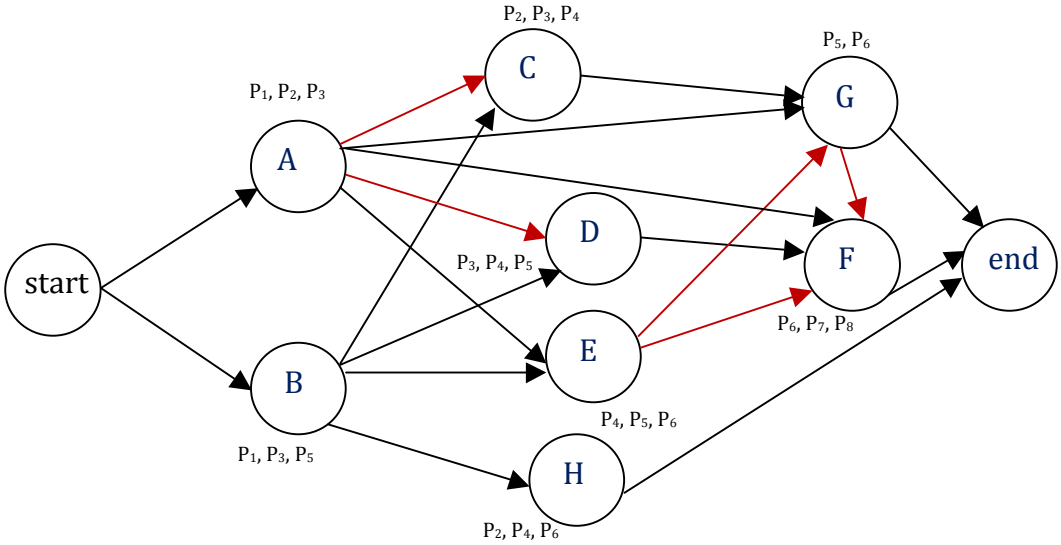


Figure 3.5 An isotopic-cluster graph

3.1.3 Assign weights by using score function

To avoid incorrectly removing peaks of fragment ions with low intensity as noisy peaks, four non-intensity features and one intensity feature of fragment ions are used to assess each possible isotopic cluster. Considering the relationships between adjacent isotopic clusters provided by isotopic-cluster graphs, a score function, which is a linear combination of five features, is used to assign weights to the edges of the isotopic-cluster graphs. To describe these five features, eight variables are defined:

$$diff1(x,y) = x - y$$
$$diff2(x,y) = x - (y + M_H)/2$$
$$diff3(x,y) = x - (y + 2 * M_H)/3$$
$$diff4(x,y) = x - (2 * y + M_H)/3$$
$$sum1(x,y) = x + y$$
$$sum2(x,y) = x + (y + M_H)/2$$
$$sum3(x,y) = x + (y + 2 * M_H)/3$$
$$sum4(x,y) = x + (2 * y + M_H)/3$$

where x is the m/z value of one peak and y is the m/z value of another peak in the rest of the spectrum, respectively; M_H is the mass of a hydrogen atom. We use diff1 and sum1 with two fragment ions represented by x and y having the same charge state (z = 1, 2, 3); diff2 and sum2 considers that the fragment ion represented by x is doubly charged and that represented by y is singly charged; diff3 and sum3 considers that the fragment ion represented by x is triply charged and that represented by y is singly charged; diff4 and sum4 considers that fragment ion represented by x is triply charged and that represented by y is doubly charged.

Four non-intensity properties of fragment ions which rely on the fragmentation technique, CID, are used to assess the possible isotopic clusters.

The first non-intensity feature (F_1') is the number of peaks y whose mass differences with a given peak x approximate the residue mass of one of the twenty amino acids. For example, if x is one of peaks in an isotopic cluster with m/z value 100, peaks with the m/z value 171.0788 or 256.1875 in a spectrum are collected as y since the relationship between their m/z values follows one of the formulas below. The differences (171.0788-100=71.0788, 256.1875-100=156.1875) are equal to the residue mass of alanine and arginine, respectively.

$$
\begin{aligned}
F_1' = |\{ y \mid & abs(diff1(x,y)) = M_{aa} + \theta \ \ or \\
& abs(diff1(x,y)) = M_{aa}/2 + \theta \ \ or \\
& abs(diff1(x,y)) = M_{aa}/3 + \theta \ or \\
& abs(diff2(x,y)) = M_{aa}/2 + \theta \ \ or \\
& abs(diff2(y,x)) = M_{aa}/2 + \theta \ \ or \\
& abs(diff3(x,y)) = M_{aa}/3 + \theta \ \ or \\
& abs(diff3(y,x)) = M_{aa}/3 + \theta \ \ or \\
& abs(diff4(x,y)) = M_{aa}/3 + \theta \ \ or \\
& abs(diff4(y,x)) = M_{aa}/3 + \theta \} |
\end{aligned}
\tag{3.1}
$$

where abs is the absolute value function; M_{aa} is the residue mass of one of twenty amino acids; |•| is the cardinality of a set; x is one of the peaks in this isotopic cluster, and y can be any peak in the rest of a spectrum; θ is the mass error tolerance. Wong et al. [40] used ±0.3 Da as the fragment ion mass error tolerance. Pan et al. [41] used ±0.5 Da as fragment ion tolerance for LTQ-Orbitrap tandem mass spectra. The space ($1/z \approx 0.3333$, $z=3$) between adjacent isotopic peaks of the triply charged fragment ion is the smallest one in the three kinds of charged fragment ions. In order to avoid counting the same fragment ion repeatedly, in this study we set the fragment ion mass tolerance as ±0.3 Da ($<1/z \approx 0.3333$, $z=3$).

The second non-intensity feature (F_2') considers that the side chains of some amino acid residues of fragment ions can lose a water molecule (H_2O) or an ammonia molecule (NH_3). The number of peaks y whose mass differences with x approximate the mass of a water molecule (H_2O) or an ammonia molecule (NH_3) is collected.

$$
\begin{aligned}
F_2' = |\{y \mid & abs(diff1(x,y)) = M_{H_2O} \ or \ M_{NH_3} + \theta \ or \\
& abs(diff1(x,y)) = M_{H_2O}/2 \ or \ M_{NH_3}/2 + \theta \ or \\
& abs(diff1(x,y)) = M_{H_2O}/3 \ or \ M_{NH_3}/3 + \theta \ or \\
& abs(diff2(x,y)) = M_{H_2O}/2 \ or \ M_{NH_3}/2 + \theta \ or \\
& abs(diff2(y,x)) = M_{H_2O}/2 \ or \ M_{NH_3}/2 + \theta \ or \\
& abs(diff3(x,y)) = M_{H_2O}/3 \ or \ M_{NH_3}/3 + \theta \ or \\
& abs(diff3(y,x)) = M_{H_2O}/3 \ or \ M_{NH_3}/3 + \theta \ or \\
& abs(diff4(x,y)) = M_{H_2O}/3 \ or \ M_{NH_3}/3 + \theta \ or \\
& abs(diff4(y,x)) = M_{H_2O}/3 \ or \ M_{NH_3}/3 + \theta\}|
\end{aligned}
\tag{3.2}
$$

where M_{H_2O} denotes the mass of a water molecule and M_{NH_3} gives the mass of an ammonia molecule; x is one of the peaks in this isotopic cluster, and y can be any peak in the rest of a spectrum; the error tolerance θ is ±0.3 Da.

The third non-intensity feature (F_3') considers two supportive ions a-ions and z-ions which can be used to indicate the existence of the corresponding b-ions and y-ions. The number of peaks representing these kinds of supportive ions is collected.

$$
\begin{aligned}
F_3' = |\{y \mid & abs(diff1(x,y)) = M_{CO} \ or \ M_{NH} + \theta \ or \\
& abs(diff1(x,y)) = M_{CO}/2 \ or \ M_{NH}/2 + \theta \ or \\
& abs(diff1(x,y)) = M_{CO}/3 \ or \ M_{NH}/3 + \theta \ or \\
& abs(diff2(x,y)) = M_{CO}/2 \ or \ M_{NH}/2 + \theta \ or \\
& abs(diff2(y,x)) = M_{CO}/2 \ or \ M_{NH}/2 + \theta \ or \\
& abs(diff3(x,y)) = M_{CO}/3 \ or \ M_{NH}/3 + \theta \ or \\
& abs(diff3(y,x)) = M_{CO}/3 \ or \ M_{NH}/3 + \theta \ or \\
& abs(diff4(x,y)) = M_{CO}/3 \ or \ M_{NH}/3 + \theta \ or \\
& abs(diff4(y,x)) = M_{CO}/3 \ or \ M_{NH}/3 + \theta\}|
\end{aligned}
\tag{3.3}
$$

where the mass of -CO is denoted by M_{CO} and the mass of -NH is denoted by M_{NH}; x is one of the peaks in this isotopic cluster, and y can be any peak in the rest of a spectrum; The error tolerance θ is ±0.3 Da.

The fourth non-intensity feature (F_4') is the number of peaks y representing fragment ions that complement with fragment ion represented by x.

$$
\begin{aligned}
F_4' = |\{ y \mid & sum1(x,y) = M + 2i + 2 \times M_H + \theta \ \ or \\
& sum1(x,y) = (M + 2i)/2 + 2 \times M_H + \theta \ \ or \\
& sum1(x,y) = (M + 2i)/3 + 2 \times M_H + \theta \ \ or \\
& sum2(x,y) = (M + 2i)/2 + 2 \times M_H + \theta \ \ or \\
& sum2(y,x) = (M + 2i)/2 + 2 \times M_H + \theta \ \ or \\
& sum3(x,y) = (M + 2i)/3 + 2 \times M_H + \theta \ \ or \\
& sum3(y,x) = (M + 2i)/3 + 2 \times M_H + \theta \ \ or \\
& sum4(x,y) = (M + 2i)/3 + 2 \times M_H + \theta \ \ or \\
& sum4(y,x) = (M + 2i)/3 + 2 \times M_H + \theta \} |
\end{aligned}
\tag{3.4}
$$

where i (0, …, 3) is the position of peak x in its isotopic cluster; M is the mass of the neutral precursor ion, and M_H is the mass of a hydrogen atom. x is one of the peaks in this isotopic cluster, and y can be any peak in the rest of a spectrum. The error tolerance θ is ±0.3 Da.

The intensity feature (F_5') determines if the experimental isotopic distribution of one possible isotopic cluster matches with a theoretical isotopic distribution considering the relationships between adjacent isotopic clusters in the graph.

Based on the natural abundance of the composition elements in one ion, the theoretical isotopic distribution of this ion can be predicted. However, the fragment ion represented by one isotopic cluster is unknown in a tandem mass spectrum. Thus, the theoretical isotopic distribution cannot be predicted precisely. Three special cases of the composition of peptide fragment ions are used to estimate the maximum, the mean and the minimum of the theoretical isotopic pattern: one is composed of all phenylalanine C_9H_9NO [5]; one is composed of an updated version of averigine $C_{4.949}H_{7.833}O_{1.473}N_{1.361}S_{0.038}$ [42]; one consists of all aspartic acid $C_4H_5NO_3$ [5]. Assume that the mass of a particular molecule is known as M, and then the number of phenylalanine units, averigine units and aspartic acid

units of this molecule can be calculated by M/the mass of phenylalanine residue, M/ the mass of averigine residue and M/the mass of aspartic residue. Then, the element composition of this molecule can be acquired. For example, the number of carbon is (9*M)/the mass of phenylalanine residue. The number of carbon calculated by phenylalanine is the largest of twenty amino acids; the one calculated by averigine is the average; the one calculated by aspartic is the smallest. In fact, that's why these special cases are used. Further, the relative natural abundance of isotopes of each element C, H, N and O is already known. Based on the information above, the maximum, mean and minimum theoretical isotopic distribution of an ion with a particular mass can be predicted.

$$F_5' = \left\{ \begin{array}{l} y' \mid \dfrac{\min\left(\left|E_i - (T_{\min})_i\right|, \left|E_i - (T_{\max})_i\right|\right)}{(T_{mean})_i} \le threshold \ \ or \\ \dfrac{\min\left(\left|(E_i - (T'_{mean})_2) - (T_{\min})_1\right|, \left|(E_i - (T'_{mean})_2) - (T_{\max})_1\right|\right)}{(T_{mean})_1} \le threshold \end{array} \right\} \quad (3.5)$$

where the first formula is for an isotopic cluster that has no shared peaks with others; the second formula is for an isotopic clusters that has shared peaks with others. E_i is the experimental intensity of peak *i*; $(T_{\min})_i$ is the minimum theoretical intensity of peak *i*, $(T_{\max})_i$ is the maximum theoretical intensity of peak *i*; $(T_{mean})_i$ is the mean theoretical intensity of peak *i*; $(T'_{mean})_i$ is the mean theoretical intensity of the other isotopic cluster which overlaps with this isotopic cluster; *i* (1, …,4) is the order of peak x in this isotopic cluster. The group of correct isotopic clusters in training dataset, which will be described in Section 4.1.1, is used to set the threshold. The experimental isotopic distributions of these isotopic clusters are compared with their corresponding theoretical isotopic distributions. The threshold is set as 0.3. Here in F_5', *y'* and *x'* belongs to the same

assumed isotopic cluster. *x'* is the first peak of the isotopic cluster, and y' is the rest of this isotopic cluster.

To assign the weights of the edges in the graph, those five features above are combined in a score function as follows,

$$score = \omega_1 \times F_1 + \omega_2 \times F_2 + \omega_3 \times F_3 + \omega_4 \times F_4 + \omega_5 \times F_5 \quad (3.6)$$

where F_i (i=1, … , 5) is the value of each feature, ω_i (i=1, … , 5) are the coefficients which were estimated using linear discriminative analysis (LDA) method of Lin *et al.* [43]. The training dataset for the estimation of coefficients are described in Section 4.1.1. We get $\omega_1 = \omega_2 = \omega_3 = 0.1$; $\omega_4 = 0.5$; $\omega_5 = 0.8$.

The possible isotopic cluster at the head of each vertex in an isotopic-cluster graph (Figure 3.6) will be assessed by the score function according to the relationship with the one at the tail of this edge. Under different relationships, the scores of the same isotopic cluster at the head of each edge are different. And the weight for each edge is calculated from the score assigned to the isotopic cluster at the head of this edge. If the head of the edge is the ending vertex of a graph, the weight of this edge is assigned as zero. The larger the weight between two connected isotopic clusters, the more reliable the assumed relationship between them is.

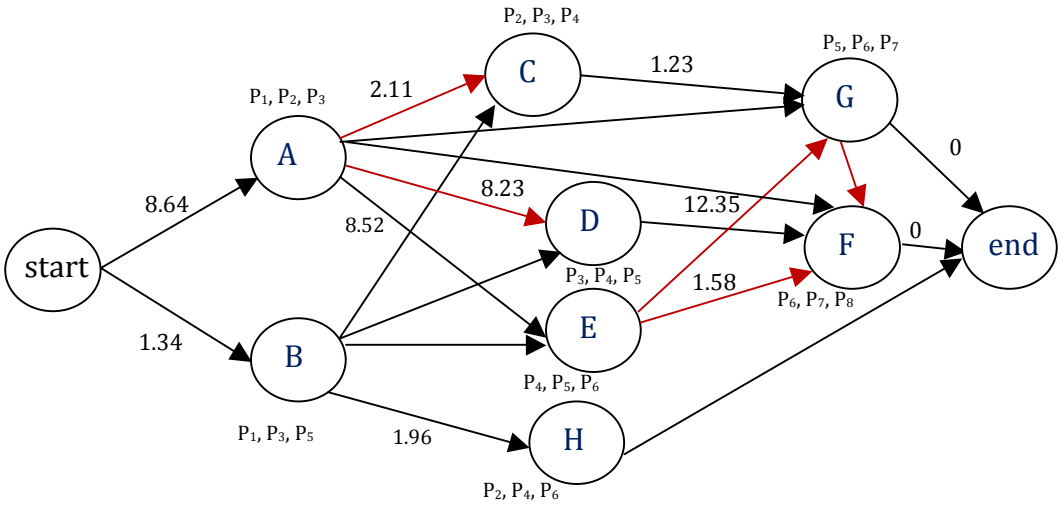


Figure 3.6 An isotopic-cluster graph with assigned weights

3.1.4 Search paths

A path in a graph is defined as a sequence of vertices connected by edges [41]. The score of a path in the isotopic-cluster graph is the sum of the weights of all edges of the path. The higher the total score of one path, the more reliably the isotopic cluster is detected. The paths with the highest score in an isotopic-cluster graph are those that cover edges with the high weights. The isotopic clusters of fragment ions are determined by searching for the optimal paths in the isotopic-cluster graphs. To identify the isotopic clusters, dynamic programming is adopted to find the path with the highest score in each isotopic-cluster graph. We wrote function the **GetBestPath** (**A, B**) to find the highest score path. **A** is an n*n matrix that shows whether any two vertices connect or not in a graph G. $\mathbf{M_{pq}}$ is the element at position (p, q) of matrix **A**. If two vertices are connected, $\mathbf{M_{pq}}$ is 1; otherwise, $\mathbf{M_{pq}}$=0. n is the number of the vertices in each graph. **B** is an n*n weighted matrix for graph G. Below is the pseudo-code for **GetBestPath**:

GetBestPath (**A, B**)

1. TP (G) is the topological ordering of graph G, denoted as ($\mathbf{v_1, v_2, \dots, v_n}$).
2. **for j=1:n**

if j=1, wst (v_j)=0; wst is an array which is defined to be the highest score of v_j;

else

while i=(1,…, j-1) & A (v_i, v_j)=1

wst (v_j)= max{wst (v_i)+B(v_i, v_j)};

label (v_j)= the vertex which maximizes wst (v_j); label is an array storing the vertex which maximizes wst(vj);

end

end

3. Output the highest score according to **wst**;
4. **While label (v_j) ≠ start** Output the highest score path according to **label**.

Push label (v_j) onto S; S is a stack.

5. **Pop (S);**

3.1.5 Determine the monoisotopic peaks

The isotopic clusters represented by the vertices of the highest score path for each graph are considered as identified. When the highest score path for each graph of a spectrum is obtained, we combine the identified isotopic clusters from each graph together.

For each identified isotopic cluster, the first peak is considered as the monoisotopic peak. And the charge state of the fragment ion represented by this isotopic cluster is estimated

based on the spaces (1/z) between the adjacent isotopic peaks. The monoisotopic mass is calculated by multiplying m/z of this monoisotopic peak with the charge.

CHAPTER 4

EXPERIMENTAL TESTS ON THE IMPROVED METHOD

4.1 Experimental Datasets

4.1.1 Training Dataset

To estimate the weights of each feature in the score function, a training dataset is constructed based on dataset of Ding *et al.* [32]. They took samples from *Escherichia coli*, digested them with trypsin, and analyzed the digest by μLC-MS/MS on a ThermoFinnigan Orbitrap LTQ mass spectrometer, yielding a total of 112329 mass spectra. Of them, 1208 high-confidence peptide-spectrum matches generated by existing algorithms [34-36] are selected for the training dataset. The thresholds for selecting the high-confidence peptide-spectrum matches are set so as to yield a FDR of 1%. The charge range of spectra is from 1 to 2 while the mass range of spectra is from 0 to 2000 Da. The training dataset consists of two groups: one group with incorrect isotopic clusters and the other group with correct isotopic clusters. Since the theoretical peptide sequences of those 1208 spectra are known, we use Peptide Fragmentation Modeller [44] to generate the theoretical fragment ions for each spectrum. Meantime, the 1208 spectra generated a list of isotopic clusters for each spectrum. Then MS-Deconv's outputs are compared with the corresponding theoretical spectra. The matched isotopic clusters are grouped as correct isotopic clusters. The rest of possible isotopic clusters of the original spectra are grouped as incorrect isotopic clusters. The weights of the score function (3.6) are estimated with this training dataset by linear discriminative analysis (LDA) method of Lin *et al.* [43].

4.1.2 Testing Datasets

To evaluate the performance of our deisotoping method, we used three LTQ-Orbitrap tandem mass spectral datasets:

(1) MS/MS dataset A [32], which is the same as the experimental dataset in Chapter 2, is used to test this improved deisotoping method.

(2) MS/MS dataset B [41] in FT2 format (a common format for mass spectrometry data) was derived from R. palustris CGA010 strain consisting of 3273 bottom-up spectra. This dataset was analyzed with a two-dimensional liquid chromatography-tandem mass spectrometry analysis (2D LC-MS/MS). Peptides eluted from the microcapillary columns were electrosprayed into an LTQ-Orbitrap mass spectrometer (ThermoFisher Scientific, San Jose, CA). The RAW format outputs of LTQ-Orbitrap mass spectrometer were converted to FT2 format. The charge range of spectra is from 1 to 3. The mass range of spectra is from 600 to 4000 Da. Our deisotoping method is compatible with the MGF file and YADA software can deal with the MS2 file. Thus we wrote two MATLAB scripts to convert the testing dataset from FT2 format to individual MGF and MS2 files, respectively.

(3) MS/MS dataset C in mzXML format (another common format for mass spectrometry data), which was provided by Pacific Northwest National Laboratory, was acquired from the organism called Shewanella oneidensis MR-1. There are 1597 bottom-up spectra from LTQ-Orbitrap mass spectrometer (ThermoFisher Scientific, San Jose, CA) in the dataset. The charge range of spectra is from 1 to 3. The mass range of spectra is from 1000 to 3000 Da. We used Trans-proteomic pipeline [45] to convert mzXML files to MGF files, which our deisotoping method is compatible with, and MS2 files which YADA software can process.

4.2 Results and Discussions

4.2.1 Performance on the testing data set A

In this section, we compare our improved deisotoping method with two pieces of software, YADA and MS-Deconv. This evaluation has two aspects: a) to see if peptide

and protein identification improves based on the number of interpreted spectra and the score of interpreted spectra by Mascot [37]; b) to see the accuracy of the determination of fragment ions.

4.2.1.1 Identification of peptides and proteins

To assess the performance of peptide and protein identification, an on-line Mascot search was performed to interpret the result sets produced by our deisotoping method, YADA and MS-Deconv. Before Mascot searching, we wrote two MATLAB scripts to convert YADA's output from an MS2 file to an MGF file, and convert MS-Deconv's output from an ENV file to an MGF file, respectively. The parameters for the Mascot search were set as given in Table 2.1. In this study, peptides are considered to be interpreted by Mascot searching engine within the false discovery rate (FDR) of 1%.

The more peptides and proteins interpreted by Mascot after being processed, the better the effect of the deisotoping method. Therefore, we used the number of interpreted peptides and proteins to assess the performance. The search results in Table 4.1 show that 281, 273, and 259 peptides are interpreted while a total of 196, 181, and 172 proteins are identified after processing by our method, YADA and MS-Deconv, respectively.

Table 4.1. Numbers of peptides and proteins identified by Mascot from dataset A (1208 spectra) processed by our method, YADA and MS-Deconv.

	Data processed by MS-Deconv	Data processed by YADA	Data processed by our method
proteins	172	181	196
peptides	259	273	281

The higher the Mascot score, the higher reliability the peptide and protein identifications are. To ensure fairness, the Mascot score comparisons with the same search parameters (Table 2.1) are based on 129 co-assigned proteins (Figure 4.1) and 172 co-assigned

peptides (Figure 4.2) from data processed by three methods. Here, “co-assigned peptides and proteins” mean that the interpreted peptides and proteins by Mascot from data processed by three methods in common. Also, in order to get objective conclusions, we use a Kruskal-Wallis statistical test to decide whether the medians of the Mascot scores from the data processed by our method, YADA, MS-Deconv are significantly different. The Kruskal-Wallis is a non-parametric method that can compare the medians of two or more samples without the assumption of normal distribution [46]. In these two figures, the Mascot scores from YADA and MS-Deconv processed data are sorted by the increasing Mascot scores from our processed data.

From Figure 4.1, although the Mascot scores of a few proteins from data processed by YADA and MS-Deconv are greater than those from our method, the median Mascot score of interpreted proteins from data produced by our method is increased by 4.3% and 7.4% over those from data produced by YADA and MS-Deconv, respectively. By using Kruskal-Wallis tests, the difference (p-value=0.5733>0.05) between the median Mascot score of the interpreted proteins from YADA processed data and the median obtained after using our method is not significant. However, the median Mascot scores of the interpreted proteins after our method and those after MS-Deconv are significantly different (p-value=$2.893*10^{-5}$<0.05).

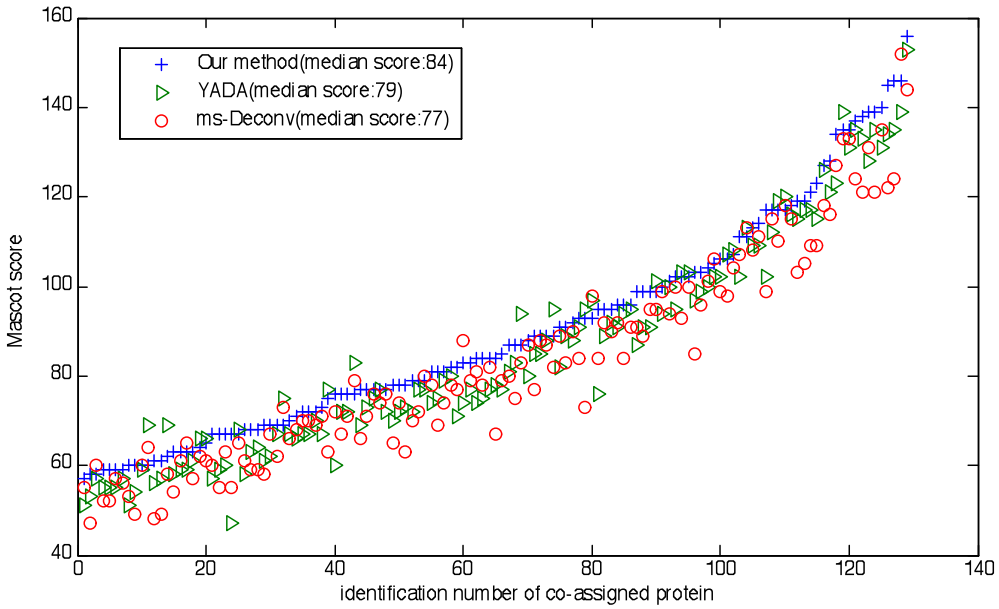


Figure 4.1 The Mascot scores of 129 proteins which are co-assigned after processing by our method (blue), by YADA (green) and by MS-Deconv (red).

From Figure 4.2, although the Mascot score of a few peptides is greater after processing by YADA and MS-Deconv than our method, the median Mascot score of the interpreted peptides after processing by our method has 6.3% and 9.1% improvement over those processed data by YADA and MS-Deconv, respectively. From the results of the Kruskal-Wallis test, the differences (p-value=0.1783>0.05) between the median Mascot scores of the interpreted peptides from YADA processed data and those after our method are not significant. However, the differences (p-value=$4.107*10^{-5}$<0.05) between the median Mascot scores of the interpreted peptides after processing by our method and those after MS-Deconv are significant.

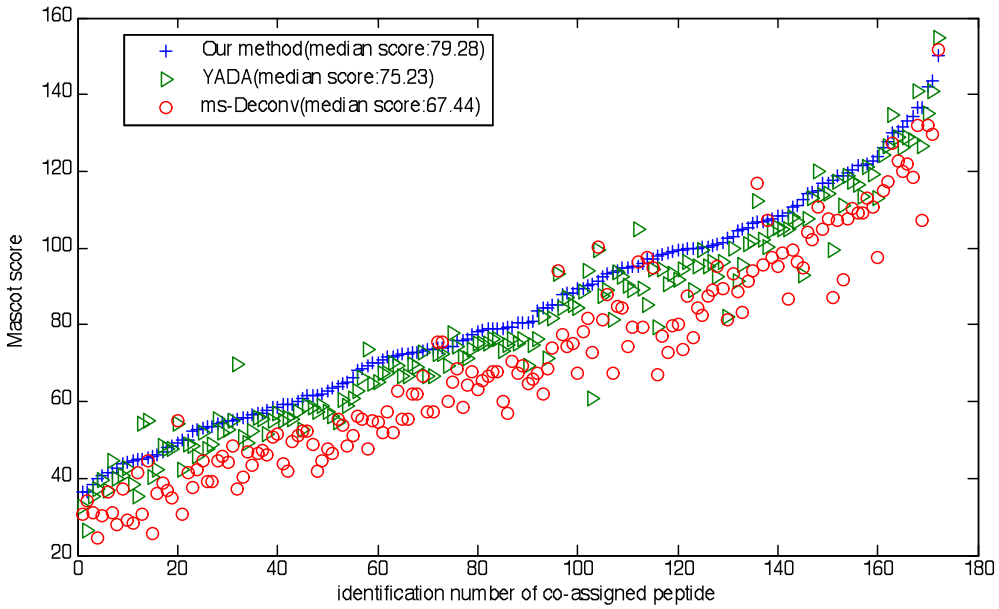


Figure 4.2 The Mascot scores on 172 peptides which are co-assigned after processing by our method (blue), by YADA (green) and by MS-Deconv (red).

Overall, though the median Mascot scores of interpreted peptides and proteins from the data processed by our method are not significantly higher than those after processing by YADA in terms of Kruskal-Wallis tests, the number of peptides and proteins interpreted after using our method (see Table 4.1) is larger than after using YADA. To some extent, our method has a better effect on the Mascot search than YADA. Moreover, not only are the numbers of peptides and proteins interpreted after our method larger than those after using MS-Deconv, but also the median Mascot scores of interpreted peptides and proteins from the data processed by our method are significantly higher than those processed by MS-Deconv. Those results indicate that the Mascot search on the data processed by our method is more reliable than the search after using MS-Deconv, and that our method performs significantly better than MS-Deconv according to a Mascot search.

4.2.1.2 Determination of monoisotopic peaks

The more real monoisotopic peaks detected by the deisotoping method, the more important information about fragment ions is obtained and the greater the accuracy of

peptide identification. To compare the performance of the real monoisotopic mass determination on the data produced by our method, YADA and MS-Deconv, F-score analysis is used.

Based on each known theoretical peptide sequence of 1208 spectra, Peptide Fragmentation Modeller [44] generated a list of theoretical fragment ions, including a, b, c, x, y, z and neutral ions. After that, a spectrum produced by our method, by YADA and by MS-Deconv was compared with its corresponding theoretical spectrum. If the difference between a peak in each experimental spectrum and a peak in its corresponding theoretical spectrum is within a given error tolerance, the peak in the experimental spectrum is regarded as a true positive (TP), and otherwise it is regarded as a false positive (FP). If the difference between a peak in a theoretical spectrum and any peak in its corresponding experimental spectrum is beyond a given error tolerance, the peak in the theoretical spectrum is regarded as a false negative (FN). We use the F-score to investigate the performance of our method, YADA and MS-Deconv. The F-score is computed by considering both the precision and the recall:

$$F = 2 * \frac{precision * recall}{precision + recall} \tag{4.1}$$

where precision is defined as TP/ (TP+FP) and recall, also called sensitivity, is defined as TP/ (TP+FN).

A series of mass error tolerances ranging from 0 to 1 Da (step 0.1 Da) were utilized for comparing an experimental spectrum with a theoretical spectrum. With different mass error tolerances, we got F-score curves shown in Figure 4.3 for three methods.

For fairness, the calculated F-scores were compared on 172 co-assigned spectra of our method's outputs, YADA's outputs and MS-Deconv's outputs. It can be observed from Figure 4.3 that under different mass error tolerances almost all F-scores from our outputs

are greater than those from YADA's outputs and MS-Deconv's outputs. It suggests that our method is more accurate than YADA and MS-Deconv in the detection of real monoisotopic peaks.

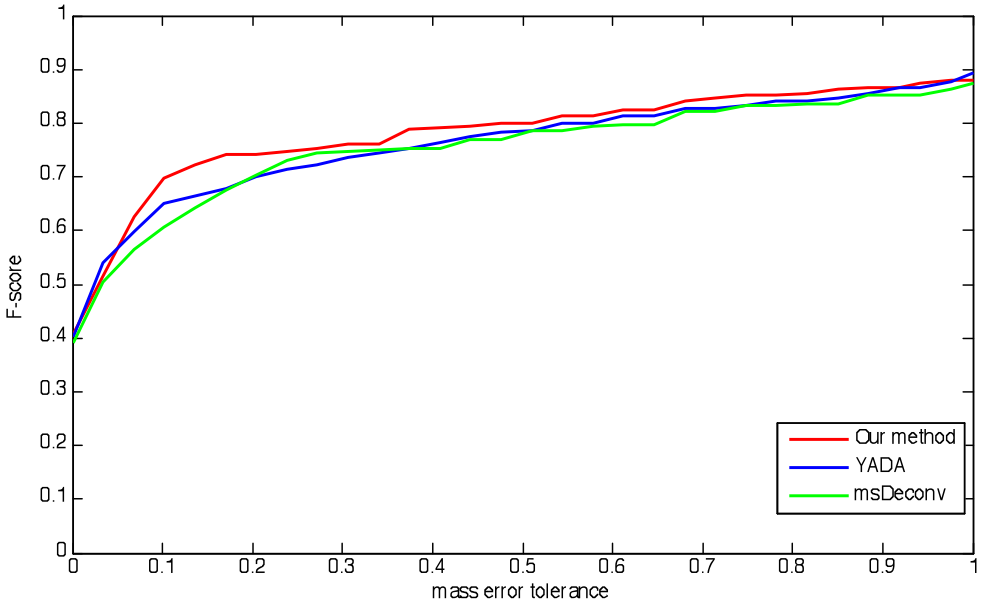


Figure 4.3 The F-scores of 172 co-assigned spectra from our method's outputs (red line), YADA's outputs (blue line) and MS-Deconv's outputs (green line).

4.2.2 Performance on the testing data sets

In this section, to investigate the performance of our method further and get more evidence, it was compared with two pieces of software, YADA and MS-Deconv, on two additional testing data sets using the same methodology as in the last section. Dataset B was used in the comparison of our deisotoping method and only the YADA software due to the format; dataset C was used to compare our method, YADA software and MS-Deconv software.

4.2.2.1 Identification of peptides and proteins

To investigate the performance of peptide and protein identification, an on-line Mascot search was employed to interpret the raw MS/MS dataset, the processed data by YADA,

the processed data by MS-Deconv and that by our deisotoping method. The searching parameters are set as specified in Table 2.1.

The effect of the deisotoping method is indicated by the number of peptides and proteins interpreted by Mascot. Table 4.2.a) shows the number of the interpreted peptides and proteins in raw data (dataset B), after processing by YADA and after our method. From this table, we can see that the number of interpreted proteins increased by 22.2% (= (143-117)/117) for the data processed by YADA, and 35.9% (= (159-117)/117 for our method. It also shows that our method can improve the number of identified peptides by 20.3% (= (231-192)/192) compared to YADA, and 40.9% (= (231-164)/164) compared to the raw data. Both increasings resulting from our method are greater than those from applying YADA. Moreover, Table 4.2.b) shows that 232, 227 and 213 peptides are interpreted while a total of 106, 105 and 101 proteins are identified from the same spectra dataset C processed by our method, YADA and MS-Deconv, respectively. The number of identified proteins by using our method is 2.2% higher than that by using YADA, and 8.9% higher than that by using MS-Deconv. The number of identified peptides by using our method increases 1.0% over that by using YADA, and 5.0% over that by using MS-Deconv. The increased rates of identified proteins and peptides by using these deisotoping methods are much lower for dataset C than those for dataset B. The reason may be that the quality of the spectra from dataset C differs from that from dataset B. The overlapping cases or the cases of fragment ions with low intensities are less frequent in dataset C than in dataset B.

Table 4.2. Numbers of peptides and proteins identified by Mascot searching from the testing data sets B and C.

a). The raw data (dataset B) and processed data by our method and YADA.

	Raw data	Data processed by YADA	Data processed by our method
proteins	117	143	159
peptides	164	192	231

b). The dataset C processed by our method, YADA software and MS-Deconv software.

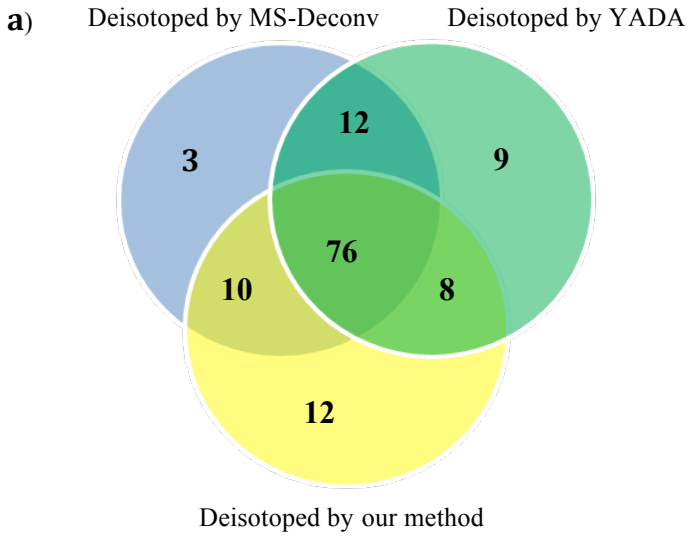
	Data processed by MS-Deconv	Data processed by YADA	Data processed by our method
proteins	101	105	106
peptides	213	227	232

In addition, Figure 4.4 shows the comparison of identified proteins and peptides from the raw data (dataset B), deisotoped data by our method and by YADA. From Figure 4.4 a), 79.7% (= (92+22)/(92+22+23+6)) of the interpreted proteins from the data processed by YADA, and 84.6% (= (92+7)/ (92+7+12+6)) for the raw data, are also identified from data processed by our method. Moreover, 23.9% (= 38/159) newly identified proteins only come from the data processed by our method. Figure 4.4 b) shows that 72.4 % (= (113+26)/ (11+113+42+26)) interpreted peptides from the processed data by YADA, and 86.0 % (= (28+113)/ (28+113+12+11)) for the raw data, are also identified from the data processed by our method. 27.7% (= 64/231) are only identified after our method. In addition, the figure shows that the number of proteins (23) and peptides (42) identified only after processing by YADA. This means that more peptides and proteins are identified by Mascot from the data processed by our method than that from the raw data and the data processed by YADA. Furthermore, Figure 4.5 a) and b) show that the number of new

peptides (56) and new proteins (12) only identified after processing dataset C by our method is larger than those from the dataset C after YADA or MS-Deconv.

In a word, the number of peptides and proteins interpreted by our method is larger than that by YADA and MS-Deconv. The results above indicate that our method has a better positive effect on the Mascot search than YADA and MS-Deconv.

Figure 4.4 Comparison of identified proteins a) and peptides b) from the raw data(dataset B), deisotoped data by our method and by YADA.



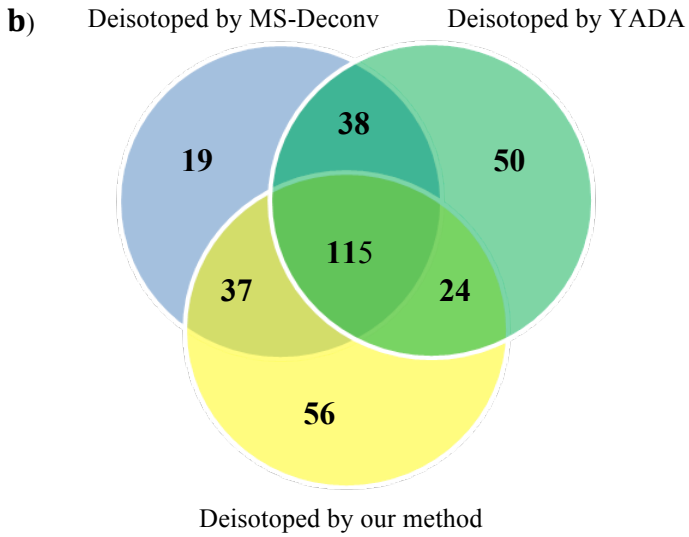


Figure 4.5 Comparison of identified proteins a) and peptides b) from the data (dataset C) deisotoped by our method, YADA and MS-Deconv.

The reliability of the peptide and protein identifications is assessed based on the Mascot score. To ensure the fairness, the Mascot score comparison is performed on the co-assigned proteins and peptides with the same parameters for Mascot searching. Also, to make the conclusions more objective, the statistic method Kruskal-Wallis tests were used

to decide whether the medians of the Mascot scores between different processed groups are significantly different.

For dataset B, Mascot scores are compared on the co-assigned proteins and peptides from the raw data and data produced by our method and YADA. Figure 4.6 shows the Mascot scores of 92 overlapped proteins from raw data of dataset B and two processed data. Compared with raw data, the median Mascot scores of the interpreted proteins from YADA processed data and from data processed by our method are increased by 31.09% (= (78-59.5)/59.5) and 56.30% (= (93-59.5)/59.5), respectively. The Kruskal-Wallis test shows that the differences (p-value=$1.479*10^{-8}$<0.05) between the median Mascot scores of the interpreted proteins from the raw data and those after processing by our method are significant. It shows that the differences (p-value=0.00022<0.05) between the median Mascot scores of the interpreted proteins from the raw data and those from YADA processed data are significant as well.

The results above indicate that the reliability of protein identification increases by applying both YADA and our method. However, our method performs better than YADA by 19.2% (= (93-78)/78). There is a significant difference (p-value=0.0310<0.05) between the median Mascot score of the interpreted proteins from YADA processed data and from data processed by our method by using Kruskal-Wallis test.

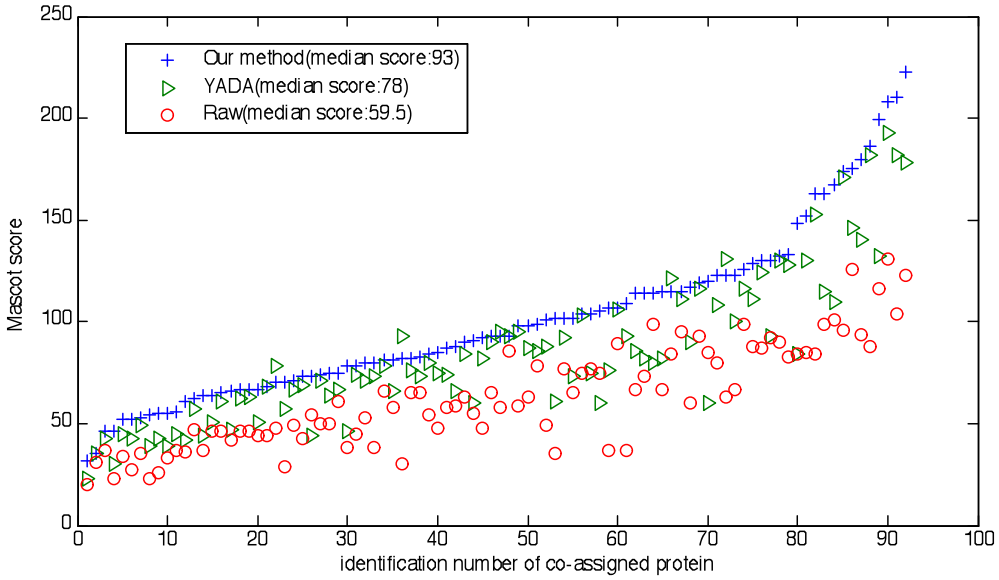


Figure 4.6 The Mascot scores of 92 proteins which are co-assigned by raw dataset B (red), data after processing by YADA (green) and by our method (blue).

The Mascot scores of 113 co-assigned peptides from raw data (dataset B) and two processed data are shown in Figure 4.7. As can be seen in the figure, the spots from both YADA and our method are higher than the spots representing the Mascot score of the raw data. The median Mascot scores of the interpreted peptides from after YADA and after our method are increased by 24.9% (= (67.06-53.67)/53.67) and 46.8% (= (78.77-53.67)/53.67), respectively, over that of the raw data. By using Kruskal-Wallis test, the differences (p-value=$1.205*10^{-11}$<0.05) between the median Mascot scores of the interpreted peptides from the raw data and those from data processed by our method are significant. The test shows that the differences (p-value=$5.404*10^{-5}$<0.05) between the median Mascot scores of the interpreted peptides from the raw data and those after YADA processing are significant as well.

Furthermore, our method has 17.5% (= (78.77-67.06)/67.06) improvement over YADA. Kruskal-Wallis test suggests there are significant differences (p-value=0.002265<0.05)

between the median Mascot scores of the interpreted peptides from YADA processed data and data processed by our method.

From the results above, the Mascot searches for both peptide and protein identifications on the data processed by our method is more reliable than those on the raw data and data processed by YADA.

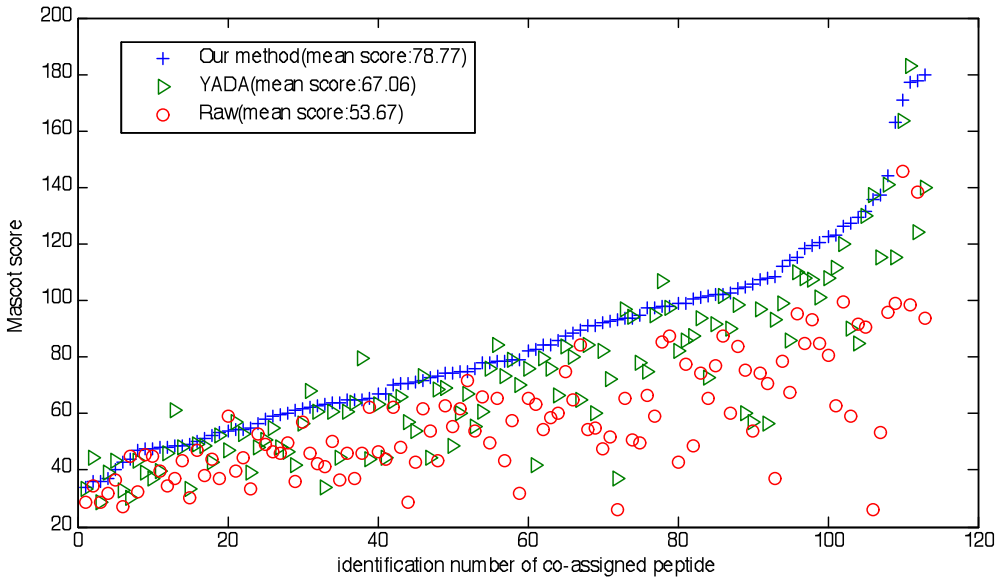


Figure 4.7 The Mascot scores of 113 peptides which are co-assigned by raw dataset B (red), data processed by YADA (green) and by our method (blue).

For dataset C, the Mascot scores are compared on the co-assigned proteins and peptides from data after processing by our method, by YADA, and by MS-Deconv. All Mascot search are performed with the same parameters.

Figure 4.8 shows the Mascot scores of 76 co-assigned proteins from processed data by our method, YADA and MS-Deconv. Although the Mascot scores of a few proteins from processed data by YADA and MS-Deconv are greater than those from our method, the median Mascot score of interpreted proteins from the processed data by our method is increased by 17.5% and 21.3% over those from processed data by YADA and MS-

Deconv, respectively. The result of Kruskal-Wallis test indicates that the differences (p-value=0.0367<0.05) between the median Mascot scores of the interpreted proteins after our method and those after processing by YADA are significant. Also, the differences (p-value=0.02578<0.05) between the median Mascot scores of the interpreted proteins from data processed by our method and those from data processed by MS-Deconv are significant as well.

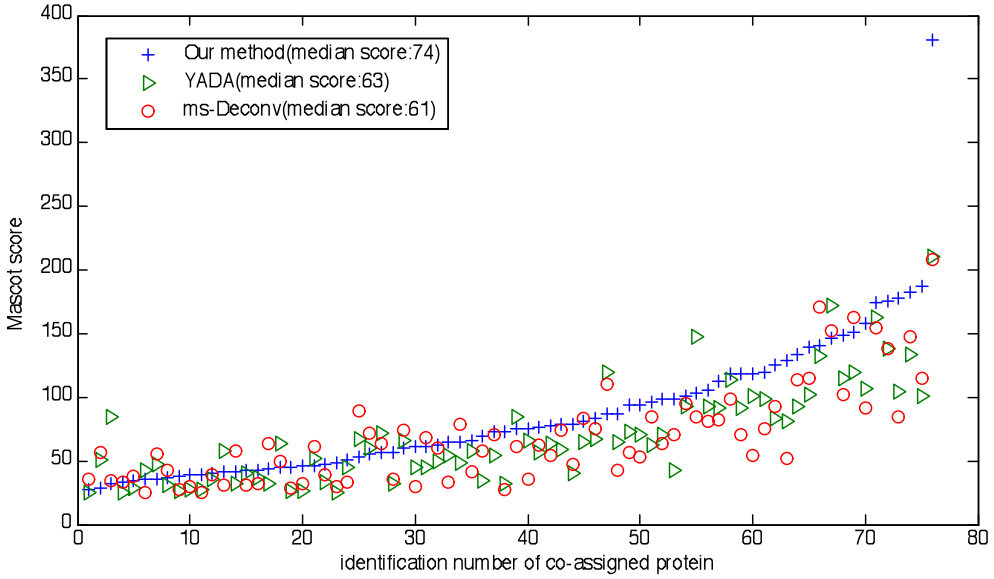


Figure 4.8 The Mascot scores of 76 proteins which are co-assigned after processing by our method (blue), by YADA (green) and by MS-Deconv (red).

Figure 4.9 shows the Mascot scores of 115 co-assigned peptides from processed data by our method, YADA and MS-Deconv. Although the Mascot scores of a few peptides from data produced by YADA and MS-Deconv is greater than those from our method, the median Mascot score of the interpreted peptides of the data produced by our method has 9.2% and 21.1% improvement over those for data produced by YADA and MS-Deconv, respectively. From the results of Kruskal-Wallis tests, the differences (p-value=0.03928<0.05) between the median Mascot scores of the interpreted peptides from YADA processed data and those from data produced by our method are significant.

Moreover, the differences (p-value=0.001125<0.05) between the median Mascot scores of the interpreted peptides from data produced by our method and those from MS-Deconv processed data are significant as well.

The higher the Mascot scores, the more reliable the peptide and protein identifications are. From the results above, the Mascot search for both peptide and protein identification from the data produced by our method is more reliable than those from YADA and by MS-Deconv.

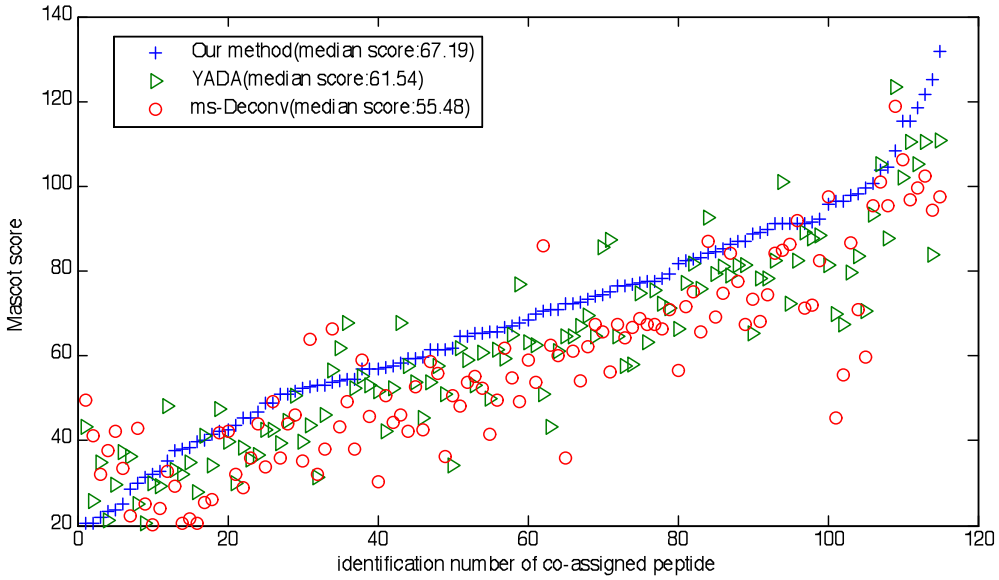


Figure 4.9 The Mascot scores of 115 peptides which are co-assigned after processing by our method (blue), by YADA (green) and by MS-Deconv (red) from original dataset C.

Moreover, in order to assess the effect of deisotoping on the speed of the Mascot analysis, the Mascot search time (in seconds) was recorded based on the elapsed time. For the dataset B, the Mascot search time of the raw data is around 121s; the Mascot search time of the data processed by YADA software is reduced to 75s; after being processed by our deisotoping method, the Mascot search time is decreased to 69s. For the dataset C, the Mascot search time of the raw data is around 104s; both the Mascot search time of the data processed by our method and YADA is reduced to around 82s; the Mascot search

time of data from MS-Deconv is around 90s. The results above illustrate that our method cannot only make peptide and protein identifications more reliable, but also reduce the Mascot searching time by providing Mascot search engine with shorter lists of monoisotopic masses.

Table 4.3. Mascot search time of the testing data sets B and C.

a). The raw data (dataset B) and processed data by our method and YADA.

	Raw data	Data processed by YADA	Data processed by our method
Time (s)	121	75	69

b). The dataset C processed by our method, YADA software and MS-Deconv software.

	Raw data	Data processed by MS-Deconv	Data processed by YADA	Data processed by our method
Time (s)	104	90	82	82

4.2.2.2 Determination of monoisotopic peaks

To evaluate the performance for the determination of real monoisotopic masses, we analyzed the true positives and the false positives.

Ⅰ. Testing dataset B (3273 spectra)

We firstly generated the theoretical peptide sequences by a de novo sequencing method called PEAKS [47], which can find the best peptide sequences from MS/MS spectra. The output of this software is a list of peptide sequences that can possibly generate the MS/MS spectra. Peptide sequences with high confidence scores are usually the correct peptide sequences. Of PEAKS' output, 2363 theoretical peptide sequences whose average local confidences are larger than 60% were selected. Then, based on each theoretical peptide sequence, Peptide Fragmentation Modeller [44] generated a list of theoretical fragment

ions, including a, b, c, x, y, z and neutral ions. After that, the spectra from our output and YADA's output were compared with each corresponding theoretical spectrum. The criteria for the determination of true positives and false positives is the same as that in Section 4.2.1.2. A series of mass error tolerances ranging from 0 to 1 Da (step 0.1 Da) were selected while comparing experimental spectra with theoretical spectra.

We used the F-score (formula 4.1) to evaluate the performance of our deisotoping method and YADA. For fairness, the calculated F-scores (shown in Figure 4.10) were compared on 139 spectra co-assigned by YADA's outputs and our method's outputs. It can be observed from Figure 4.10 that under different mass error tolerances almost all F-scores from our outputs are greater than those from YADA's outputs. This indicates that our method is more accurate than YADA in the detection of real monoisotopic peaks.

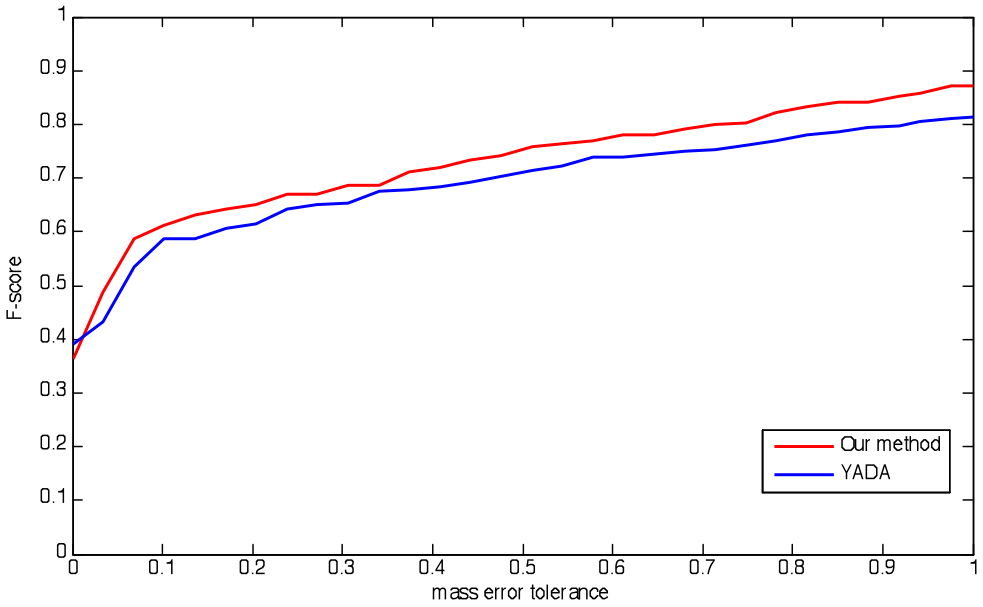


Figure 4.10 The F-scores of 139 co-assigned spectra from our method's outputs (red line) and YADA's outputs (blue line).

Ⅱ. Testing dataset C (1597 spectra)

The theoretical peptide sequences of testing dataset C were generated by using PEAKS. Of PEAKS' output, 1084 theoretical peptide sequences whose average local confidences are larger than 60% were selected. Based on each theoretical peptide sequence, Peptide Fragmentation Modeller generated a list of theoretical fragment ions, including a, b, c, x, y, z and neutral ions. After that, the spectra from our output, YADA's output and MS-Deconv's output were compared with each corresponding theoretical spectrum with an error tolerance 0.1 Da.

Box-and-whisker plots were drew to compare the performances of our method, YADA and MS-Deconv in the determination of monoisotopic peaks. Boxplots can show not only the distribution of the data, but also can suggest whether or not the differences between two median values are significant. If the notches of two boxes do not overlap, it can indicate the medians are significant different (at the 5% level). Figure 4.11 shows the distribution of the true positives of 115 spectra co-assigned from data processed by our method, YADA, and MS-Deconv. From this figure, we can see that the notch of our method's box doesn't overlap with the notches of the other two boxes. It indicates that the box of our method has the significantly different median from those two boxes. We can conclude that the spectra produced by our method have more true positives than those from data produced by YADA, and MS-Deconv.

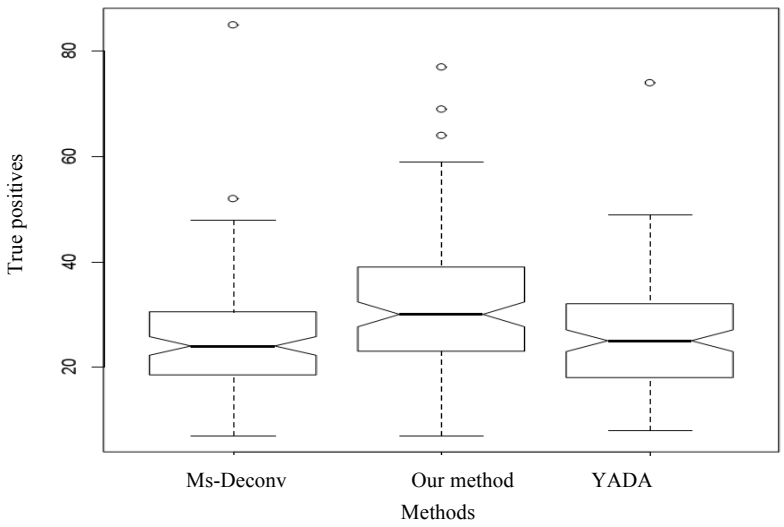


Figure 4.11 The boxplot graphic of the number of true positives from 115 spectra which are co-assigned by data processed by our method, data processed by YADA and by MS-Deconv.

The distribution of the false positives of 115 co-assigned spectra from data processed by our method, YADA, and MS-Deconv is shown in Figure 4.12. The notch for the box of our method might overlap with the notch of the box for YADA. However, it definitely does not overlap with the notch of the box for MS-Deconv. Thus, the median of the box for our method may be significant from that of the box for YADA. Though the median of the box for our method is not significantly different from the medians of the box for MS-Deconv, some spectra from data processed by our method has fewer false positives than those from data processed by MS-Deconv.

From the results above, 115 co-assigned spectra from data processed by our method yield more true positives than those from data processed by YADA and MS-Deconv. In addition, the numbers of false positives from 115 co-assigned spectra from data processed by our method are not more than those from data processed by the comparison software.

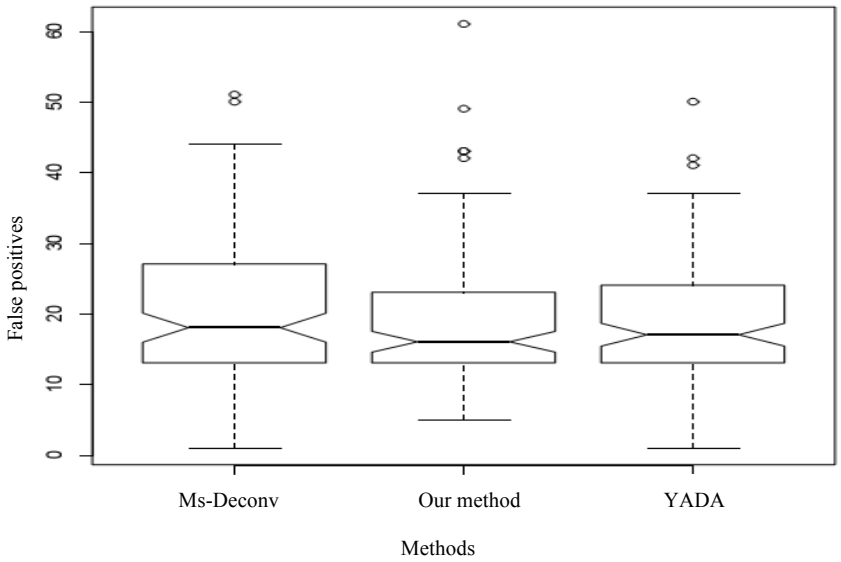


Figure 4.12 The boxplot of the number of false positives for 115 spectra which are co-assigned by data processed by our method, data processed by YADA and by MS-Deconv.

CHAPTER 5

CONCLUSIONS AND FUTURE WORK

5.1 Conclusions

This thesis firstly shows a preliminary study of a feature-based deisotoping method for tandem mass spectra. In order to solve the problems of this method, we have presented an improved deisotoping algorithm for bottom-up spectra to increase the accuracy of monoisotopic mass determination of fragment ions. Since peaks of fragment ions with low intensities can be removed as noisy peaks by accident, in this study we have explored non-intensity properties of fragment ions in tandem mass spectra: mass relationships, relative intensity ratios of adjacent isotopic peaks, the relationships between the isotopic distribution of fragment ions and that of precursor ions and so on. Moreover, the existence of overlapping cases could cause the missing of monoisotopic peaks in overlapping cases. We solved this problem by analyzing the relationships between possible isotopic clusters.

In Chapters 3 and 4, the improved algorithm takes overlapping cases into account by firstly constructing isotopic-cluster graphs which describe the relationship between possible isotopic clusters. Based on the assumed relationships in the graphs, all possible isotopic clusters are evaluated by a score function which combines non-intensity and intensity features of fragment ions. According to the relationships between isotopic clusters provided by the isotopic cluster graphs, each candidate isotopic cluster is given a score based on the score function. Dynamic programming is adopted to find the path with the highest score as the optimal arrangement of isotopic clusters with the highest reliability.

The experimental results from three data sets have indeed indicated that our improved method performs better in deisotoping compared with YADA and MS-Deconv software in three aspects:1) a larger number of interpreted proteins and peptides from the dataset

processed by our deisotoping method indicates our method has better effect on Mascot searching; 2) the peptide and protein identifications from the data produced by our method is more reliable than those from the other two kinds of software based on higher median Mascot scores. The conclusions based on the results of Kruskal-Wallis tests have statistical significance; and 3) the F-scores of our method are greater than those after using the other two kinds of software.

5.2 Future work

By applying our algorithm, we improved the accuracy of the peptide and protein identification. However, there are still some problems deserving more attention. Firstly, the deisotoping process is still time-consuming because of some naïve strategies used in the path-search algorithm. For each spectrum, there could be a lot of isotopic-cluster graphs and we had to search each isotopic-cluster graph for the path with the highest score. Thus, if the path is too long (>30 nodes), it would take too much time. In this case, we need to try other algorithms to improve the searching speed instead of dynamic programming.

Secondly, while using isotopic-cluster graphs to describe the relationships between possible isotopic clusters, only predominant overlapping cases were taken into account. This may result in missing fragment ions in very complex overlapping cases.

To improve the performance of our feature-based deisotoping method, more features of isotopic distributions can be defined and incorporated into the score function. Many features of isotopic distribution have been proposed in existing literature such as charge states; distances between peaks within a potential distribution; the shapes; and the number of peaks in the potential distribution; the relative intensities of the two highest peaks in the distribution; and so on. Such features can be evaluated to determine if they are relevant to the identification of isotopic clusters. Moreover, we will test our deisotoping method on more mass spectral datasets.

REFERENCES

[1] C. Dass. Fundamentals of contemporary mass spectrometry. *John Wiley & Sons,* 2007.

[2] J.T. Watson, O.D. Sparkman. Introduction to mass spectrometry. *John Wiley & Sons*, 2007.

[3] J.S. Davies, G.C. Barrett. Amino acids, peptides, and proteins. *Royal Society of Chemistry,* 2003.

[4] N. Mujezinovic, G. Raidl, J.R. Hutchins, J. Peters, K. Mechtler and F. Eisenhaber. Cleaning of raw peptide MS/MS spectra: Improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. *Proteomics*, 6 (19): 5117-5131, 2006.

[5] K. Park, J. Y. Yoon, S. Lee, E. Paek, H. Park, H. Jung, and S.W. Lee. Isotopic peak intensity ratio based algorithm for determination of isotopic clusters and monoisotopic masses of polypeptides from high-resolution mass spectrometric data. *Anal. Chem*, 80 (19): 7294-7303, 2008.

[6] D. Valkenborg, I. Jansen, and T.Burzykowski. A model-based method for the prediction of the isotopic distribution of peptides. *J Am Soc Mass Spectrom,* 19: 703-712, 2008.

[7] I. Eidhammer, K. Flikka, L. Martens, S.O. Mikalsen. Computational methods for mass spectrometry proteomics. *John Wiley & Sons,* 2008.

[8] S. Gay, P.A. Binz, D.F. Hochstrasser, R.D. Appel. Modeling peptides mass fingerprinting data using the atomic composition of peptides. *Electrophoresis*, 20: 3527-3524, 1999.

[9] J.F. Zhang, S. He, J.J. Cai, X.J. Cao, R.X. Sum, Y. Fu, R. Zeng and W. Gao.

Processing of tandem mass spectrometric data cased on decision tree classification. *Geno. Prot. Bioinfo*, 3 (4): 231-237, 2005.

[10] R.M. Smith. Understanding mass spectra: a basic approach. *John Wiley & Sons,* 2004.

[11] J.F. Zhang, S. He, C.X. Ling, X.J. Cao, R. Zeng and W. Gao. PeakSelect: preprocessing tandem mass spectra for better peptide identification. *Rapid Commun. Mass Spectrom*, 22 (8): 1203-1212, 2008.

[12] Y. Sun, J. Zhang, U.B. Neto, E.R. Dougherty. BPDA−A Bayesian peptide detection algorithm for mass spectrometry. *BMC Bioinformatics*, 11: 490, 2010.

[13] J.F. Zhang, D. Xu, W. Gao, G.H. Lin and S. He. Isotope pattern vector based tandem mass spectral data calibration for improved peptide and protein identification. *Rapid Commun. Mass Spectrom.* 23 (21): 3448-3456, 2009.

[14] D.M. Horn, R.A. Zubarev, F.W. McLafferty. Automated reduction and interpretation of high resolution electrospray mass spectra of large molecules. *J Am Soc Mass Spectrom*, 11(4): 320-332, 2000.

[15] P. Kaur, PB. O'Connor. Algorithms for automatic interpretation of high resolution mass spectra. *J Am Soc Mass Spectrom,* 17 (3): 459–468, 2006.

[16] K. Noy, D. Fasulo. Improved model-based, platform-independent feature extraction for mass spectrometry. *Bioinformatics*, 23 (19): 2528-2535, 2007.

[17] N. Jaitly, A. Mayampurath, K. Littlefield, J.N. Adkins, G.A. Anderson, R.D. Smith. Decon2LS: An open-source software package for automated processing and visualization of high resolution mass spectrometry data. *BMC Bioinformatics,* 10: 87, 2009.

[18] M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, T.O. Schulz, A. Zerck, K. Reinert, O. Kohlbacher. OpenMS - An open- source software framework for mass spectrometry. *BMC Bioinformatics*, 9: 163, 2008.

[19] C. Masselon, L.P. Tolic, S.W. Lee, L.J. Li, A. Gordon, A. Richard, H. Richard, D. Smith. Identification of tryptic peptides from large databases using multiplexed tandem mass spectrometry: simulations and experimental results. *Proteomics*, 3(7): 1279-1286, 2003.

[20] P.C. Carvalho, T. Xu, X.M. Han, D. Cociorva, V.C. Barbosa1 and J.R. Yates. YADA: a tool for taking the most out of high-resolution spectra. *Bioinformatics,* 25(20), 2734-2736, 2009.

[21] J.A. Yergey. A general approach to calculating isotopic distributions for mass spectrometry. *Int J Mass Spectrom Ion Phys,* 52(2-3): 337-349, 1983.

[22] A.L. Rockwood, S.L. Van Orden, R. Smith. Rapid calculation of isotope distributions. *Anal Chem*, 67(15): 2699-2704, 1995.

[23] M.W. Senko, S.C. Beu, F.W. McLafferty. Automated assignment of charge states from resolved isotopic peaks for multiply charged ions. *J Am Soc Mass Spectrom,* 6(1): 52-56, 1995.

[24] J. Samuelsson, D. Dalevi, F. Levander, T. Rögnvaldsson. Modular, scriptable and automated analysis tools for high-throughput peptide mass fingerprinting. *Bioinformatics,* 20(18): 3628-3635, 2004.

[25] B.Y. Renard , M. Kirchner, H. Steen, J.A. Steen, and F. A. Hamprecht. NITPICK: peak identification for mass spectrometry data. *BMC Bioinformatics,* 9: 355, 2008.

[26] P. Du, R.H. Angeletti. Automatic deconvolution of isotope-resolved mass spectra using variable selection and quantized peptide mass distribution. *Anal Chem*, 78: 3385-

3392, 2006.

[27] R. Tibshirani. Regression shrinkage and selection via the lasso. *J Am Soc Mass Spectrom*, 58 (1): 267-288, 1996.

[28] J. Zhang, H. Wang, A. Suffredini, D. Gonzales, E. Gonzales, Y. Huang, X. Zhou. Bayesian peak detection for pro-TOF MS MALDI data. *Proc of IEEE International Conference on Acoustics, Speech and Signal Processing*, 661-664, 2008.

[29] X. Li, E.C. Yi, C.J. Kemp, H. Zhang, R. Aebersold. A software suite for the generation and comparison of peptide arrays from sets of data collected by liquid chromatography-mass spectrometry. *Mol Cell Proteom*, 4 (9): 1328-1340, 2005.

[30] S. McIlwain, D. Page, E.L. Huttlin and M.R. Sussman. Using dynamic programming to create isotopic distribution maps from mass spectra. *Bioinformatics*, 23 (13): 328-336, 2007.

[31] X.W. Liu, Y. Inbar, P.C. Dorrestein, C. Wynne, N. Edwards, P. Souda, J.P. Whitelegge, V. Bafna and P.A. Pevzner. Deconvolution and Database Search of Complex Tandem Mass Spectra of Intact Proteins: A Combinatorial Approach. *Mol Cell Proteomics,* 9 (12): *2772-2782,* 2010.

[32] J.R. Ding, J.H. Shi, G.G. Poirier, and F.X. Wu. A novel approach to denoising ion trap tandem mass spectra. *Proteome Sci.* 7: 9, 2009.

[33] A.A. Klammer, S.M. Reynolds, J.A. Bilmes, M.J. MacCoss, W.S. Noble. Modeling peptide fragmentation with dynamic Bayesian networks for peptide identification. *Bioinformatics,* 24 (13): 348-356, 2008.

[34] J.K. Eng, A.L. McCormack, J.R. Yates. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J Am Soc Mass Spectrom.* 5 (11): 976-989, 1994.

[35] L.Y. Geer, S.P. Markey, J.A. Kowalak, L. Wanger, M. Xu, D.M. Maynard, X. Yang, W. Shi, S.H. Bryant. Open mass spectrometry search algorithm. *J. Proteome Res*. 3: 958-964, 2004.

[36] N. Zhang, R. Aebersold, B. Schwikowski. ProbID: a probabilistic algorithm to identify peptides through sequence database searching using tandem mass spectral data. *Proteomics*, 2(10): 1406-1412, 2002.

[37] D.N. Perkins, D.J. Pappin, D.M. Creasy, J.S. Cottrell. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis,* 20 (18): 3551-3567, 1999.

[38] F.X. Wu, P. Gagné, A. Droit, G.G. Poirier. Quality assessment of peptide tandem mass spectra. *BMC Bioinformatics*, 9: S13, 2008.

[39] N.C. Jones, P. Pevzner. An introduction to bioinformatics algorithms. *MIT Press Cambridge*, 2004.

[40] J. Wong, M. Sullivan, H. Cartwright, G. Cagney. msmsEval: tandem mass spectral quality assignment for high-throughput proteomics. *BMC Bioinformatics*, 8: 51, 2007.

[41] C. Pan, B.H Park, W.H. McDonald, P.A. Carey, J.F. Banfield, N.C. VerBerkmoes, R. L. Hettich, N.F. Samatova. A high-throughput de novo sequencing approach for shotgun proteomics using high-resolution tandem mass spectrometry. *BMC Bioinformatics,* 11:118, 2010.

[42] N. Barbarini, P. Magni. Accurate peak list extraction from proteomic mass spectra for identification and profiling studies. *BMC Bioinformatics,* 11: 518, 2010.

[43] W. Lin, F.X. Wu, J. Shi, J. Ding, and W.J. Zhang. An adaptive weight approach to denoising ion trap tandem mass spectra. *BIBMW,* 89-94, 2010.

[44] http://omics.pnl.gov/software/PeptideFragmentationModeller.php. August 2, 2011.

[45] http://tools.proteomecenter.org/wiki/index.php?title=Software:TPP. May 17, 2012

[46] P. Dalgaard. Introductory statistics with R. *Springer*, 2008.

[47] B. Ma, K.Z. Zhang, C. Hendrie, C.Z. Liang, M. Li, A.D. Kirby and G. Lajoie. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom.* 17 (20): 2337-2342, 2003.